

POISSON AND INDEPENDENT PROCESS APPROXIMATION FOR RANDOM COMBINATORIAL STRUCTURES WITH A GIVEN NUMBER OF COMPONENTS, AND NEAR-UNIVERSAL BEHAVIOR FOR LOW RANK ASSEMBLIES

RICHARD ARRATIA AND STEPHEN DESALVO

ABSTRACT. We give a general framework for approximations to combinatorial assemblies, especially suitable to the situation where the number k of components is specified, in addition to the overall size n . This involves a Poisson process, which, with the appropriate choice of parameter, may be viewed as an extension of saddlepoint approximation.

We illustrate the use of this by analyzing the component structure when the rank and size are specified, and the rank, $r := n - k$, is small relative to n . There is near-universal behavior, in the sense that apart from cases where the exponential generating function has radius of convergence zero, for $\ell = 1, 2, \dots$, when $r \asymp n^\alpha$ for fixed $\alpha \in (\frac{\ell}{\ell+1}, \frac{\ell+1}{\ell+2})$, the size L_1 of the largest component converges in probability to $\ell + 2$. Further, when $r \sim tn^{\ell/(\ell+1)}$ for a positive integer ℓ , and $t \in (0, \infty)$, $\mathbb{P}(L_1 \in \{\ell + 1, \ell + 2\}) \rightarrow 1$, with the choice governed by a Poisson limit distribution for the number of components of size $\ell + 2$. This was previously observed, for the case $\ell = 1$ and the special cases of permutations and set partitions, using Chen-Stein approximations for the indicators of attacks and alignments, when rooks are placed randomly on a triangular board. The case $\ell = 1$ is especially delicate, and was not handled by previous saddlepoint approximations.

CONTENTS

1. Introduction	2
1.1. Overview of this paper	2
1.2. Connections with enumerative combinatorics	3
1.3. Decomposable combinatorial structures	5
1.4. The conditioning relation	6
1.5. Three classes of examples: selections, multisets, and assemblies	7
1.6. The finite case, T_n , versus the infinite case, T	10
2. Choosing uniformly from the $p(n, k)$ objects with k components	11
2.1. Poisson process, conditional on having k arrivals	12
2.2. Two versions of the k -Boltzmann Sampler	14
3. Application: low rank structures	16

Date: July 4, 2016.

3.1. Low rank assemblies	17
3.2. The critical regime for having components of size 3	18
3.3. A completely effective version of Theorem 3.3	28
4. Acknowledgements	32
References	32

1. INTRODUCTION

1.1. Overview of this paper. There are good independent process approximations for random decomposable combinatorial structures, including the broad classes of assemblies, multisets, and selections; see [9]. In particular for assemblies, though not for the the other two, the approximating independent variables optimally come from a Poisson process. This specific structure leads to an especially effective representation of assemblies with a given number of components, as summarized in (23) – (25).

The plan of this paper is as follows: We will first review conditional independent process descriptions of general combinatorial objects. Then, specializing to assemblies, we will describe the additional Poisson structure, and the ensuing saddlepoint analysis. We conclude with a specific application, a universal limit law for low rank assemblies. The conclusions of this low rank limit law do *not* apply to multisets or selections: while polynomials over finite fields are very close to random permutations in many properties of the joint distribution of the sizes of the parts, the *conditional* distributions, conditioned on the existence of very many parts, are quite different.

Here is an overview of the universal low rank limit for assemblies. First, Theorem 3.3 handles the key critical case, picking an assembly uniformly at random from the $p(n, k)$ instances of size n having exactly k components, in a regime where, for some fixed $t > 0$, n and k tend to infinity together with $n - k \sim t\sqrt{n}$. ([29, Theorem 3] gives some information about a very different large deviation regime, including cases where $k \sim tn^\alpha$ for fixed $\alpha \in (0, 1/2)$.) The behavior is that the largest component has, with high probability, size 2 or 3; the number of components of size 3 has a limit Poisson distribution. This case cannot be handled by traditional saddlepoint analysis, where Gaussian behavior plays the crucial role.

Two special assemblies are set partitions, and permutations, so that $p(n, k)$ is a Stirling number of the second or first kind. For these two cases, the Poissonian behavior when $n - k \sim t\sqrt{n}$ was discovered in [6] and [5], and the limit behavior was derived from Chen-Stein analysis of combinatorial bijections, involving $n - k$ rooks placed on a triangular board of size n . For most aspects of component behavior, permutations and set partitions have very different behavior, for example a random permutation tends to have around $\log n$ cycles but a random set partition tends to have around $n/\log n$ blocks. The *only* thing permutations and set partitions have in common is

the structure of an *assembly*; the Poissonian low rank limit law does *not* apply to multisets, such as random polynomials over a finite field.

Theorem 3.3 has an asymptotic error bound of the form $O_t(\log^2 n/\sqrt{n})$. It is essential to be aware that for a specific instance of n and k , big O asymptotics provide no information at all. Theorem 3.19 carries out essentially the same analysis, but gives a quantitative, i.e., completely effective, error bound, as in [48]. Thus, the bound in Theorem 3.19 is a complicated but explicit function $u_M(n, k)$ of n and k , and the parameters of the assembly, with the property that under the regime with $n - k \sim t\sqrt{n}$, asymptotically we have $u_M(n, k) = O_t(\log^2 n/\sqrt{n})$ — thereby giving an alternate proof of Theorem 3.3.

We round out the analysis of low rank assemblies with theorems treating the low rank behavior on both sides of the critical regime $n - k \sim t\sqrt{n}$. On the side where $n - k$ is smaller, handling the case $n - k \rightarrow \infty$ with $(n - k)/\sqrt{n} \rightarrow 0$, the error bound in Theorem 3.13 is quantitative and asymptotically sharp. On the side where $n - k$ is larger, Theorem 3.14 handles the case where $n - k \asymp n^\alpha$ for fixed $\alpha \in (1/2, 1)$. For each $\ell = 2, 3, \dots$, there is a critical case: when $\alpha = \ell/(\ell + 1)$, that the largest component has, with high probability, size $\ell + 1$ or $\ell + 2$; the number of components of size $\ell + 2$ has a limit Poisson distribution.

1.2. Connections with enumerative combinatorics. Our results are parallel to classical enumerative combinatorics, where there has been much interest in counting the number of such restricted structures and proving certain smoothness conditions like unimodality and log-concavity. Much of the theoretical treatment centers on saddle point analysis, whereas our approach is distinctly probabilistic. Nevertheless, one sees the same quantities appear both in the classical analysis of generating functions as well as the probabilistic treatment involving conditional distributions. The goal is the same, which is to describe the internal shape of these interconnected structures, in which there are many statistics of interest, e.g., enumeration, smoothness, component sizes, limit shapes, large deviations, etc. The underlying theme is quantifying the severity of dependence between the components with respect to any desired statistic.

A motivating example is the set of integer partitions of size n into exactly k parts, for which there is an extensive history. Integer partitions are an example of a multiset, and thus are not covered in our scope. In addition, specifically for integer partitions there is a well-known bijection which allows one to instead consider the number of partitions of n into parts of size at most k , and the generating function is most obliging in this setting for asymptotic analysis [23, 51, 52]. Erdős and Lehner looked at $k \sim \frac{\sqrt{n}}{2c} \log n$, which governs the size of the largest part in a uniformly chosen integer partition of size n . For k small, i.e., $k = o(\sqrt{n})$, there is one asymptotic formula for $p(n, k)$ that holds uniformly [51]. For larger k , if one is content with formulas involving implicit parameters which change character depending

on k , then [52, Theorem 1] provides a complete answer, which includes the Hardy-Ramanujan formula as a special case; this was later proved by elementary means in [17] and through probabilistic means in [47]. One of the motivations for obtaining such enumerative formulas is presented in [52, Theorem 2], which demonstrates that the number of partitions of size n into exactly k parts is unimodal in k . Similar treatments were undertaken for partitions into distinct parts [1], and indeed unimodality and log-concavity is a traditional and lively topic in enumerative combinatorics [50, 16]. From a probabilistic point of view, such smoothness properties of a sequence correspond to central and local limit theorems, see [11, 12, 13].

A closer parallel, and indeed a special case which motivated the present document, is the asymptotic analysis surrounding Stirling numbers of the second kind, denoted by $S(n, k)$, corresponding to the basic combinatorial assembly, set partitions. There is an equally extensive history for Stirling numbers, see for example [38, 18, 6] and the references therein, which begins with the elementary analysis of Jordan [32] which covers $S(n, k)$ and $S(n, n - r)$ for k and r fixed. As with integer partitions, one must decide what constitutes an answer, as the seminal work of Moser and Wyman [39] essentially characterizes the asymptotic behavior for the entire range of values of k , *albeit with implicitly defined parameters*. A detailed analysis was recently carried out by Louchard [38], in which he obtained an asymptotic expansion in terms of explicit parameters, but with gaps at $r \sim t n^{\ell/(1+\ell)}$ for $t > 0$ and $\ell = 1, 2, \dots$. The gap at $\ell = 1$ was filled by the authors in [6], using a bijection involving non-attacking rooks on a lower triangular chess board. The intuition is that while saddle point analysis paints a broad brush, it is essentially capturing a central limit theorem, dominated by a normal distribution; however, at $r \sim t n^{\ell/(1+\ell)}$, the behavior is Poissonian, and so the usual analytical methods either break down or become obfuscated by the implicitly defined parameters. Stirling numbers of the first kind follow a similar tradition, see for example [32, 40, 37, 6, 18] and the references therein.

Another important aspect of asymptotic analysis is the distinction between obtaining big O error bounds, with statements such as *there exists an n_0 such that* . . . , versus quantitative bounds, which provide statements such as *for all $n \geq 26$* For example, one can prove using the first term of the Hardy-Ramanujan asymptotic formula [30, Equation (5.5)] that *there exists an n_0 such that* the number of integer partitions of n is a log-concave sequence for all $n \geq n_0$, but their analysis cannot specify the smallest value of n for which this property is guaranteed to hold¹. In order to obtain $n_0 = 26$, one would need the quantitative bounds provided by Rademacher [44] or

¹Note that one should not attempt to prove even the asymptotic log-concavity of the partition function using [30, Equation (1.41)]; see [21, Section 7] and <http://tinyurl.com/kkc6fwf>.

Lehmer [36], along with some elementary albeit tedious analysis, see [41]; see also [21].

1.3. Decomposable combinatorial structures. *Decomposable* combinatorial structures of size n are often examined with respect to the underlying integer partition of n . For a given instance of the structure of size n , having k components, the integer partition is written $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_k)$. Here λ_j is the size of the j th largest component, so that $n = \lambda_1 + \dots + \lambda_k$, and $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_k \geq 1$. Picking uniformly at random from the $p(n)$ structures of size n , the corresponding random integer partition is (L_1, L_2, \dots, L_K) , where we write $K \equiv K_n$ for the number of components, and $L_j \equiv L_j(n)$ is defined to be the size of the j th largest component, with the added provision that $L_j = 0$ if $j > K$, so that the random variable L_j is defined for all $j \in \mathbb{N}$, the set of positive integers. We write \mathbb{Z}_+ for the set of non-negative integers.

An alternate way to report the information in L_1, L_2, \dots is to consider $C_i \equiv C_i(n)$, the number of components of size i ; of course $C_i(n) := 0$ if $i > n$. The notation $C_i \equiv C_i(n)$ says that we consider both C_i and $C_i(n)$ to be the same; which notation gets used depends on whether or not one wishes to emphasize the role of the parameter n . The entire process of component counts is

$$(1) \quad \mathbf{C} \equiv \mathbf{C}(n) := (C_1(n), C_2(n), \dots, C_n(n)).$$

To review, the following two identities are trivial: $C_1 + 2C_2 + \dots = n$, $C_1 + C_2 + \dots = K$.

Independent process approximations for random combinatorial structures, conveniently abbreviated as IPARCS, would be, in the greatest generality, a choice of decomposable combinatorial structure, and independent non-negative integer-valued random variables Z_1, Z_2, \dots such that, with respect to *some* functionals $\phi : \mathbb{Z}_+^n \rightarrow \mathbb{R}$ of interest, the distribution of $\phi(\mathbf{C}(n))$ is well-approximated by the distribution of $\phi((Z_1, \dots, Z_n))$. Examples of natural functionals of interest include:

- The least common multiple of all component sizes, $\text{lcm}(L_1, \dots, L_k)$.
- The indicator, 0 or 1, of the statement that all component sizes are distinct.
- The size L_1 of the largest component.
- The difference $L_1 - L_2$, or the ratio L_2/L_1 , for the largest and second largest component.
- The number of components, K versus its approximation $Z_1 + \dots + Z_n$.
- The number of components of size at most $b(n)$, versus $Z_1 + \dots + Z_b$, for some given function $b : \mathbb{N} \rightarrow \mathbb{N}$.
- The number of components of size at least $a(n)$, versus $Z_a + \dots + Z_n$, for some given function $a : \mathbb{N} \rightarrow \mathbb{N}$.
- The number of components of size in the range $[a(n), b(n)]$.

- The *process* of all small components, $(C_1, \dots, C_{b(n)})$, versus (Z_1, \dots, Z_b) , for some given function $b : \mathbb{N} \rightarrow \mathbb{N}$.
- The *process* of all large components, $(C_{a(n)}, \dots, C_n)$, versus (Z_a, \dots, Z_n) , for some given function $a : \mathbb{N} \rightarrow \mathbb{N}$.

For each of the above functionals, there are examples of fundamental and natural combinatorial objects where a sensibly chosen independent process gives a good approximation, with respect to that functional, and there are other combinatorial examples where a sensibly chosen independent process does not give a good approximation. Good examples for the last two functionals listed above are given by Pittel [42, 43], for integer partitions and set partitions, respectively. For contrast, an example of a functional where one *never* gets a good approximation: the functional is the weighted sum of the component counts, so that $\phi(\mathbf{C}(n)) = C_1(n) + 2C_2(n) + \dots + nC_n(n)$ is the constant random variable of value n , while $\phi((Z_1, Z_2, \dots, Z_n)) = Z_1 + 2Z_2 + \dots + nZ_n$ is not constant.

1.4. The conditioning relation. In addition to having independent random variables Z_1, Z_2, \dots, Z_n which give a usable approximation to a combinatorial structure $\mathbf{C}(n)$, it is often the case, as surveyed in [9] and [3], that there is a one parameter family of distributions, indexed by $x > 0$ or x in a bounded range such as $(0, 1)$, such that for every choice of x , there is an equality of joint distributions, after conditioning on the event that $T_n = n$, where

$$(2) \quad T_n := Z_1 + 2Z_2 + \dots + nZ_n;$$

that is,

$$(3) \quad \textbf{Conditioning Relation: } \mathcal{L}(\mathbf{C}(n)) = \mathcal{L}_x((Z_1, Z_2, \dots, Z_n) \mid T_n = n).$$

When the value of the parameter x has been chosen, the above might be denoted more simply as $\mathbf{C}(n) =^d (Z_1, Z_2, \dots, Z_n) \mid T_n = n$. Implicit in the conditioning relation and the combinatorial setup is a relation, which in the case of *combinatorial assemblies* is (15), expressing $p(n)$ exactly, in terms of the normalizing constants for the random variables Z_i – these vary with x , and $\mathbb{P}_x(T_n = n)$, which also varies with x .

In the context of probability theory, the *saddle point heuristic* [19, 31, 45] is that probability approximations, such as those recognizable by the factor $1/\sqrt{2\pi\sigma^2}$ from the central limit theorem, tend to be more accurate when appropriately tilting, perhaps by the Cramer tilt (called the Esscher tilt in [31], after [24]). Thanks to (15), we can recognize saddlepoint approximations to $p(n)$, the number of size n instances of a given type of assembly, as corresponding to approximations for $\mathbb{P}_x(T_n = n)$ carried out by picking a value of the parameter x for which $\mathbb{E}_x T_n$ is near n .

The *extended* saddle point heuristic is in one sense more specific: it says that in the x -indexed family of distributions for the right side of (3), when one wants to *remove* the conditioning on $T_n = n$ and use the independent process (Z_1, Z_2, \dots, Z_n) as an approximation (with respect to various

functionals ϕ) for $\mathbf{C}(n)$, good approximations are obtained by picking the parameter x for which $\mathbb{E}_x T_n$ is equal to, or close to, its target in the conditioning, n .

The main point of the present paper is to extend this extended saddle point heuristic to combinatorial assemblies of size n and having k components, where the given size $k = k(n)$ may be far from the typical number of components of a random structure of size n . We provide a construction, (23) – (25), which forces the number of components to be a given k , and still leaves a parameter x and a sequence of independent random variables Y_1, \dots, Y_k , such that *conditional* on the event $Y_1 + \dots + Y_k = n$, we achieve *exactly* the distribution of the component structure $\mathbf{D}(n, k)$ of a random assembly of size n with k components, with all $p(n, k)$ possibilities equally likely. As a heuristic, when the parameter x is chosen so that $\mathbb{E}(Y_1 + \dots + Y_k)$ is close to n , the independent sample Y_1, \dots, Y_k is close in distribution, with respect to various functionals ϕ , to the distribution of $\mathbf{D}(n, k)$.

1.5. Three classes of examples: selections, multisets, and assemblies. We take these three main classes in order of the superficial complexity of description; in particular, assemblies come last.

Selections and multisets To begin, one assumes that there is a universe U of objects having a positive integer-valued weight, such that for $i = 1, 2, \dots$, the number m_i of objects of weight i satisfies: m_i is finite.

Selections, in this context, are simply all finite subsets of U , with the natural notion, that the weight of a set is the sum of the weights of its elements. Let $p(n)$ be the number of sets of weight n ; always, $p(0) = 1$, with the emptyset being the unique set of weight 1. The characterizing ordinary generating function for this story is

$$(4) \quad P(z) := \sum_{n \geq 0} p(n) z^n = \prod_{i \geq 1} (1 + x^i)^{m_i}.$$

The corresponding independent random variables for the conditioning relation (3) are $\text{Binomial}(m_i, x^i/(1 + x^i))$, so that the distribution of Z_i is the m_i -fold convolution of the Bernoulli($p = x^i/(1 + x^i)$) distribution on $\{0, 1\}$.

Multisets, in this context, are all finite cardinality multisubsets of U , with the natural notion, that the weight of a multiset is the sum of the weights of its elements. Let $p(n)$ be the number of multisets of weight n ; always, $p(0) = 1$, with the emptyset being the unique multiset of weight 1. The characterizing ordinary generating function for this story is

$$(5) \quad P(z) := \sum_{n \geq 0} p(n) z^n = \prod_{i \geq 1} (1 - x^i)^{-m_i}.$$

The corresponding independent random variables for the conditioning relation (3) are $\text{NegativeBinomial}(m_i, x^i)$, so that the distribution of Z_i is the m_i -fold convolution of the Geometric (starting from zero, with ratio x^i) distribution on $\{0, 1, 2, \dots\}$, that is, the distribution of G with $\mathbb{P}(G \geq k) = x^{ik}$, $k = 0, 1, 2, \dots$.

Examples of multisets and selections include

- **Integer partitions** Here $m_i = 1$ for all $i \geq 1$, and the sole object of weight i in the universe U is the integer i itself. Multisets correspond to integer partitions, with no restrictions, and our $p(n)$ is the usual p_n , satisfying the asymptotic relation

$$(6) \quad p_n \sim \frac{e^{2\sqrt{n\pi^2/6}}}{\sqrt{48n}},$$

as found, and also more effectively approximated, by Hardy, Ramanujan, Rademacher, and Lehmer, [30, 44, 35, 36]. Selections correspond to integer partitions with all parts distinct.

- **Polynomials over \mathbb{F}_q .** Here, the universe U is the set of monic irreducible polynomials over the finite field \mathbb{F}_q with q elements, with weight being degree. By unique factorization, multisets correspond to monic polynomials, and putting $p(n) = q^n$ in (5) leads to the relations $q^n = \sum_{i|n} im_i$, with $m_i \equiv m_i(q)$, which by Möbius inversion is equivalent to $im_i = \sum_{d|i} \mu(i/d)q^d$, so that

$$(7) \quad m_i(q) = \frac{1}{i} \sum_{d|i} \mu(i/d)q^d \sim \frac{q^i}{i};$$

this was known to Gauss, see [14]. The explicit formula shows that as $i \rightarrow \infty$, $m_i(q) = (q^i/i)(1 + O(q^{-i/2}))$ which is a remarkably easy and effective analog of the prime number theorem. While multisets correspond to all monic polynomials, *selections* correspond to *square-free* monic polynomials. The generating function (5) does not converge at $x = 1/q$, but $x = 1/q$ is the correct value to use, so that the independent Z_i have $\mathbb{E}_x Z_i \sim 1/i$ as $i \rightarrow \infty$, and we have a logarithmic combinatorial structure in the sense of [3], with the number of irreducible factors of a random polynomial of degree n growing like $\log n$.

- **Necklaces over an alphabet of size q .** Same as the above, without the restriction that q is a prime power; see for example [53].

Assemblies appear in 1974, called abelian partitioned composite, and recognizable via the exponential relation (13), in Foata [27]. Assemblies are called *species* by Joyal [33, 34]; they are called exponential families, with decks and hands, by Wilf [54]; they are called *uniform structures* in [53, Theorem 14.2]; and they are discussed extensively as the *SET* construction for labelled structures in [26, Section II]. Wilf [54, Section 3.18] supplies the early history: Riddell in a 1951 thesis [46, Footnote 18] has the exponential formula (13) in the context of simple graphs. Bender and Goldman in 1971 [10] and Foata and Schützenberger in 1970 [28] have the general assembly, called *prefab* in the former, and *composé partitionnelle* in the latter.

The simplest assembly is *set partition*, equivalently, arbitrary *equivalence relation*: for a set partition of size n , the set $[n] := \{1, 2, \dots, n\}$ is decomposed as a disjoint union of nonempty subsets, referred to as the *blocks* or *equivalence classes*, and the blocks are gathered as a set of blocks, rather than a list of blocks. For an integer partition of n having counts \mathbf{a} , that is, a_i parts of size i , $i = 1$ to n , so that

$$(8) \quad \mathbf{a} = (a_1, a_2, \dots, a_n) \in \mathbb{Z}_+^n, \text{ with } a_1 + 2a_2 + \dots + na_n = n,$$

the number of set partitions of *type* \mathbf{a} , that is, having a_i blocks of size i , $i = 1$ to n , is

$$(9) \quad N(n, \mathbf{a}) = n! \frac{1}{\prod_1^n a_i! (i!)^{a_i}}.$$

The general assembly is specified by a sequence of nonnegative integers m_1, m_2, \dots , which is encoded in the exponential generating function

$$(10) \quad M(z) := \sum_{i \geq 1} \frac{m_i z^i}{i!}.$$

If one doesn't explicitly state the range of summation, $i \geq 1$, then one should specify that $m_0 = 0$. An instance of this assembly, of size n , is formed in two steps: 1) pick a set partition on $[n]$, and 2) for each block of size i , *decorate* that block in one of m_i ways. Hence, the the number of M -assemblies of *type* \mathbf{a} is

$$(11) \quad N(n, \mathbf{a}) = n! \prod_1^n \frac{m_i^{a_i}}{a_i! (i!)^{a_i}}.$$

Summing over all \mathbf{a} satisfying (8) yields $p(n)$, the total number of M -assemblies of size n ; wrapping these up in an exponential generating function yields

$$(12) \quad P(z) := \sum_{n \geq 0} p(n) z^n / n!.$$

Note that we always have $p(0) = 1$ for the trivial but confusing reason that there is a unique equivalence relation on the empty set, and since this has no blocks, there is exactly one way to decorate all these blocks.

The succinct characterization of an M -assembly is the exponential relation on exponential generating functions,

$$(13) \quad P(z) = \exp(M(z)).$$

This is always a valid relation, in the sense of formal power series; as an exercise, the reader is urged to name the radius of convergence of each of the examples presented after (15).

The corresponding independent random variables for the conditioning relation (3) are Poisson, with

$$(14) \quad \mathbb{E}_x Z_i = \lambda_i \equiv \lambda_i(x) := \frac{m_i x^i}{i!},$$

valid for any $x > 0$. The identity (3) is very easily proved, by comparison with (9), and by equating the normalizing constants for $\mathbb{P}(\mathbf{C}(n) = \mathbf{a})$ and $\mathbb{P}_x((Z_1, Z_2, \dots, Z_n) = \mathbf{a})$, where \mathbf{a} as in (8) has weighted sum $a_1 + 2a_2 + \dots + na_n = n$, whence we get the *identity*

$$(15) \quad p(n) = \frac{n!}{x^n} \exp(\lambda_1(x) + \dots + \lambda_n(x)) \mathbb{P}_x(T_n = n).$$

Examples of assemblies include

- **Set partitions**, decomposed into blocks. Here, $m_i = 1$ for all $i \geq 1$, so $M(z) = e^z - 1$, and $p(n)$ is usually denoted as B_n , the n th Bell number.
- **Permutations**, decomposed into cycles. Here, $m_1 = m_2 = 1, m_3 = 2$; in general $m_i = (i-1)!$. We have $M(z) = \log(1-z)$, $p(n) = n!$, and $P(z) = 1/(1-z)$.
- **Random mappings**, i.e., arbitrary functions $f : [n] \rightarrow [n]$. Of course, $p(n) = n^n$. The components are the weakly connected components of the random mapping digraph, i.e. the directed graph exactly n edges, namely $(i, f(i))$. Here $m_1 = 1, m_2 = 2, m_3 = 17$, and it turns out that $m_i = (i-1)! \sum_{j=0}^{i-1} i^j / j!$.
- **Simple graphs**, i.e., undirected graphs with no loops and no multiple edges. Here, $p(n) = 2^{\binom{n}{2}}$, and for small i , m_i can be computed from the relation $P(z) = e^{M(z)}$. It is easy to show that as $i \rightarrow \infty$, $m_i \sim p(i)$. Hence $M(\cdot)$ has radius of convergence zero, and does *not* satisfy the hypotheses of Theorems 3.3. This assembly does not follow the behavior described by Theorems 3.3: picking uniformly from the $p(n, k)$ simple graphs on n vertices with exactly k components, for $k = n - r$ where $r = \lfloor \sqrt{n} \rfloor$, it is fairly easy to see that with probability tending to 1, there is one large component, of size $r + 1$, and the other $n - r - 1$ components are singletons.

1.6. The finite case, T_n , versus the infinite case, T . In some situations involving the conditioning relation (3) and the finite sum T_n in (2), the infinite sum

$$(16) \quad T := Z_1 + 2Z_2 + \dots$$

is more convenient to work with. The random variable T takes values in the extended nonnegative integers, $\{0, 1, 2, \dots, \infty\}$, and one of the requirements for T to be useful is that $\mathbb{P}_x(T < \infty) = 1$. When T is to be used, recalling that $C_i(n) := 0$ whenever $i > n$, the conditioning relation (3) is changed to

$$(17) \quad \mathbf{CR}: \quad \mathcal{L}((C_1(n), C_2(n), \dots)) = \mathcal{L}_x((Z_1, Z_2, \dots) \mid T = n).$$

(If one is not using the finite sum T_n , but only the infinite sum T , it would make sense to reuse the notation from (1), and define $\mathbf{C} \equiv \mathbf{C}(n) := (C_1(n), C_2(n), \dots)$, but since the purpose of this section is to clarify the similarities and differences between the two setups, we don't bother giving $(C_1(n), C_2(n), \dots)$ its own symbol.) Note that (3) is valid, in the example

where the combinatorial structure is simple graphs, described at the end of Section 1.5, even though the parameter x is necessarily outside the circle of convergence, and $\mathbb{P}_x(T = \infty) = 1$. But this is *not* an informative example for the choice T_n versus T , since for this example, the extended saddle point heuristic, described at the end of Section 1.4, does not provide useful approximation for *any* value of the parameter x .

An informative example is random permutations. The exponential generating function in (12) is $P(z) = 1/(1-z)$, so that at $z = 1$, the series diverges. Shepp and Lloyd [49] consider the conditioning relation (17) involving T , with parameter $x < 1$, so that $\mathbb{E}Z_i = x^i/i$ and $\mathbb{E}T = 1/(1-x) < \infty$, and take but $x \rightarrow 1$ to get results by applying a Tauberian theorem. In contrast, [8, Theorem 2] uses elementary analysis to get a completely effective error bound, with superexponentially fast decay, by considering $x = 1$ but using T_n , with $\mathbb{E}_x T_n = n$ at $x = 1$.

Another informative example is polynomials over \mathbb{F}_q . The ordinary generating function in (5) is $P(z) = \sum_{n \geq 0} p(n)z^n = \sum_{n \geq 0} q^n z^n = 1/(1-qz)$, and again, the most useful parameter choice is $x = 1/q$, on the boundary of the region of convergence, but where the series converges and $1 = \mathbb{P}(T = \infty)$. One can work with the choice $x = 1/q$ and the conditioning relation, as in [3], or more directly get approximations involving the Z_i , which are NegativeBinomial($m_i, x^i/(1+x^i)$), with $x = 1/q$, using inclusion-exclusion, as in [2].

2. CHOOSING UNIFORMLY FROM THE $p(n, k)$ OBJECTS WITH k COMPONENTS

For a given type of decomposable combinatorial object, let $p(n, k)$ be the number of instances of size n , having exactly k components. For M -assemblies, $p(n, k)$ can be computed from (11) by summing $N(n, \mathbf{a})$ over all \mathbf{a} with both $a_1 + 2a_2 + \dots + na_n = n$ and $a_1 + a_2 + \dots + a_n = k$. More succinctly, $p(\cdot, \cdot)$ is determined by the two variable generating function relation

$$(18) \quad \sum_{n, k \geq 0} p(n, k) \frac{z^n \theta^k}{n!} = e^{\theta M(z)},$$

which reduces to (13) by setting $\theta = 1$.

For set partitions, $p(n, k)$ is typically denoted $S(n, k)$, and is called a Stirling number of the second kind. For permutations, $p(n, k)$ is typically denoted $s(n, k)$ or $|s(n, k)|$, and is called an (unsigned) Stirling number of the first kind.

Pick uniformly from the $p(n, k)$ instances of a structure of size n with k components. Similar to (1), we write $D_i \equiv D_i(n, k)$ for the number of components of size i . The entire process of component counts is

$$(19) \quad \mathbf{D} \equiv \mathbf{D}(n, k) = (D_1(n, k), D_2(n, k), \dots, D_n(n, k)).$$

To review, the following two identities are trivial: $D_1 + 2D_2 + \cdots = n$, $D_1 + D_2 + \cdots = k$. Recall that in the similar statement following (1), we wrote $C_1 + C_2 + \cdots = K$, with the uppercase K being the random number of components of a uniformly chosen instance of size n ; here for $D_1 + D_2 + \cdots$ we have the lowercase k , which is constant (even though for applications one takes $n, k \rightarrow \infty$ together, which is usually described via $k = k(n)$.)

Trivially, picking uniformly from a finite set B (all $p(n)$ assemblies of size n) and then conditioning on landing in a given subset B_0 (all $p(n, k)$ assemblies of size n with k components) yields the uniform distribution on the subset B_0 . Hence for every n, k , the distribution of $\mathbf{D}(n, k)$ is the conditional distribution of $\mathbf{C}(n)$ given that $K = k$:

$$(20) \quad \mathbf{D}(n, k) =^d (\mathbf{C}(n) \mid K = k).$$

It then follows from (3) combined with (20) that for every $x > 0$

$$(21) \quad \text{CR: } \mathcal{L}(\mathbf{D}(n, k)) = \mathcal{L}_x((Z_1, Z_2, \dots, Z_n) \mid T_n = n \text{ and } Z_1 + \cdots + Z_n = k).$$

(This may be confusing because of the change in notation, from K to $Z_1 + \cdots + Z_n$, but it is genuinely trivial, and corresponds to the associative property of multiplication: we are biasing the distribution of (Z_1, \dots, Z_n) first by multiplying in the indicator of the event $Z_1 + 2Z_2 + \cdots + nZ_n = n$, and then multiplying in the indicator of the event $Z_1 + Z_2 + \cdots + Z_n = k$.)

Finally, as in Section 1.6, one may also work with the infinite sum T from (16), to get

$$(22) \quad \text{CR: } \mathcal{L}((D_1(n, k), D_2(n, k), \dots)) = \mathcal{L}_x((Z_1, Z_2, \dots) \mid T = n \text{ and } Z_1 + Z_2 + \cdots = k).$$

2.1. Poisson process, conditional on having k arrivals. Consider a general Poisson process on a space S , with intensity measure μ ; for simplicity of exposition, we restrict to the case $\lambda := \mu(S) < \infty$. See for example [55, Section II.37]. (A note on notation: in (14) – (15), and in this section, λ and λ_i serve as Poisson parameters, as is typical notation in standard probability texts; in contrast, in the sections highlighting combinatorial arguments, λ_i denotes the i th part of a partition λ of the integer n , as is standard notation in combinatorics texts.) The Poisson process is characterized by the requirement that for disjoint (measurable) $B_1, B_2, \dots, B_r \subset S$, with $N_i \equiv N(B_i)$ defined as the number of arrivals in B_i , one has that N_i is Poisson with parameter $\mu(B_i)$, and N_1, \dots, N_r are mutually independent. Given μ , there is a very simple construction of the desired Poisson process: Let Y be a random element of S , with distribution $(1/\lambda)\mu$, and let Y, Y_1, Y_2, \dots be i.i.d., and independent of a random variable Z , taken to be Poisson with parameter λ . Now one simply defines the (multiset) of all arrivals to be the sample of random size Z , i.e., the multiset $\{Y_1, Y_2, \dots, Y_Z\}$, so that for any measurable $B \subset S$, $N(B) = \sum_{i \geq 1} 1(Y_i \in B, i \leq Z)$. (A similar story holds in the case where μ is a sigma-finite measure on S with $\mu(S) = \infty$, but one has to

restrict to regions $R \subset S$ with $\mu(R) < \infty$, and some care must be taken to put the sigma-finite pieces together.)

The result of the above-described coupling is that, conditional on having k arrivals overall, the arrivals *are* the i.i.d. sample of size k , i.e., $\{Y_1, Y_2, \dots, Y_k\}$ considered as a multisubset of S . For the purpose of approximating the component structure of assemblies, one takes $S = [n]$ if the intention is to use (3), and $S = \mathbb{N}$ if the intention is to use (17). In either case, to write out the explicit recipe, recall (14), that $\lambda_i \equiv \lambda_i(x) := m_i x^i / i!$, and let

$$(23) \quad \mathbb{P}_x(Y = i) = \begin{cases} \frac{\lambda_i}{\lambda_1 + \dots + \lambda_n} & \text{if } i \in S = [n] \\ \frac{\lambda_i}{\lambda_1 + \lambda_2 + \dots} = \frac{\lambda_i(x)}{M(x)} & \text{if } i \in S = \mathbb{N}, \text{ and } M(x) < \infty. \end{cases}$$

Let Y, Y_1, Y_2, \dots be i.i.d., and independent of a random variable Z , taken to be Poisson with parameter λ , with $\lambda = \lambda_1 + \dots + \lambda_n$ for the case $S = [n]$, and $\lambda = \lambda_1 + \lambda_2 + \dots = M(x)$, for the case $S = \mathbb{N}$ and $M(x) < \infty$. We use these to construct, simultaneously, the Poisson process, and for each $k = 0, 1, 2, \dots$, a realization of the Poisson process conditional on having k arrivals overall.

We view the above construction as a *coupling*. Given a value $k \geq 0$, we write $N_i \equiv N_i(k)$ for the count of how many of Y_1, \dots, Y_k are equal to i . Hence $N_1 + \dots + N_n = k$ in case $S = [n]$, or $N_1 + N_2 + \dots = k$, in case $S = \mathbb{N}$. We have, for each $x > 0$

$$(N_1, \dots, N_n) =^d ((Z_1, \dots, Z_n) | Z_1 + \dots + Z_n = k), \text{ if } S = [n]$$

and

$$(N_1, N_2, \dots) =^d ((Z_1, Z_2, \dots) | Z_1 + Z_2 + \dots = k), \text{ if } S = \mathbb{N}, \text{ and } M(x) < \infty.$$

Note that the sum of the arrivals Y_j is the weighted sum of the counts, i.e., in the case $S = [n]$, always $Y_1 + \dots + Y_n = N_1(Z) + 2N_2(Z) + \dots + nN_n(Z)$. Hence by further conditioning on the weighted sum of the Z_i being equal to n , (21) and (22) imply that

$$(24) \quad \mathbf{CR}: \quad \mathcal{L}(\mathbf{D}(n, k)) = \mathcal{L}_x((N_1, N_2, \dots, N_n) \mid Y_1 + \dots + Y_n = n).$$

(25)

$$\mathbf{CR}: \quad \mathcal{L}((D_1(n, k), D_2(n, k), \dots)) = \mathcal{L}_x((N_1, N_2, \dots) \mid Y_1 + \dots + Y_n = n).$$

Each of (24) and (25) leads to an exact expression, similar to (15) for $p(n)$, but now relating $p(n, k)$ and $\mathbb{P}_x(Y_1 + \dots + Y_n = n)$. Let \mathbf{a} satisfy (8) and the additional condition that $a_1 + a_2 + \dots + a_n = k$. Then (11) implies that

$$(26) \quad \mathbb{P}(\mathbf{D}(n, k) = \mathbf{a}) = \frac{n!}{p(n, k)} \prod_1^n \frac{m_i^{a_i}}{a_i! (i!)^{a_i}}.$$

On the right side of (24) and (25), the conditional probability has denominator $\mathbb{P}_x(Y_1 + \dots + Y_k = n)$ and numerator

$$(27) \quad \mathbb{P}_x((N_1, N_2, \dots) = \mathbf{a}) = \binom{k}{a_1, a_2, \dots, a_n} \prod_1^n (\mathbb{P}_x(Y = i))^{a_i}.$$

Combining (14); the first case of (23); (24) as the statement that $\mathbb{P}(\mathbf{D}(n, k) = \mathbf{a}) = \mathbb{P}_x((N_1, N_2, \dots, N_n) = \mathbf{a}) / \mathbb{P}_x(Y_1 + \dots + Y_k = n)$; (26); (27); and cancelling common factors, we get

$$(28) \quad \forall x > 0, \quad p(n, k) = \frac{n!}{k!} \frac{(\lambda_1(x) + \dots + \lambda_n(x))^k}{x^n} \mathbb{P}_x(Y_1 + \dots + Y_k = n).$$

Using instead the second case of (23), and (25), in a similar way we get

$$(29) \quad \forall x: M(x) < \infty, \quad p(n, k) = \frac{n!}{k!} \frac{(M(x))^k}{x^n} \mathbb{P}_x(Y_1 + \dots + Y_k = n).$$

Note that for the above, so long as $a_1 + 2a_2 + \dots + na_n = n$ and $a_1 + a_2 + \dots + a_n = k$, all factors depending on \mathbf{a} cancel, leading to (28) and (29), which are identities in x , with no trace of \mathbf{a} left behind. Perhaps it will come as a surprise that there is another strategy for using estimates of $\mathbb{P}_x(Y_1 + \dots + Y_k = n)$ to give estimates for $p(n, k)$, for which the key is to name a *specific* pivotal choice of \mathbf{a} . We carry this out in Corollary 3.4, to get asymptotics for $p(n, k)$ when $n - k \sim t\sqrt{n}$.

2.2. Two versions of the k -Boltzmann Sampler. The idea of “Boltzmann sampling”, popularized by [22], is that when one wants to sample a structure of a given size n , uniformly distributed over the $p(n)$ possibilities of that size, it may be useful to ignore the requirement of getting size exactly n , and instead generate a random object of size T_n having mean around n . From our point of view, this is simply the combination of (2) and (3) while not requiring the occurrence of the conditioning event $\{T_n = n\}$, together with the extended saddle point heuristic, that the conditioning has a mild effect on many functionals of the joint distribution of (Z_1, Z_2, \dots, Z_n) when the tilting parameter x is chosen so that $\mathbb{E}_x T_n$ is close to the target n . One virtue of [22] relative to [9] is that the classes studied include more than just assemblies, multisets, and selections.

Now suppose one wants, given n and k , to sample a structure, uniformly distributed over the $p(n, k)$ possibilities of size n having exactly k components. For the special case of assemblies, the Poisson-process-inspired conditioning relations, (24) and (25), provide a very convenient analog of the Boltzmann sampler. This analog is to pick the parameter x for the distribution of Y in (23) so that $k\mathbb{E}_x Y$ is close to the target n , then take an i.i.d. sample Y_1, \dots, Y_k from this distribution, and ignore the requirement that $Y_1 + \dots + Y_k = n$. In effect, we generate a random structure of random size $Y_1 + \dots + Y_k$ by a method that *guarantees* having *exactly* k components; conditional on the event that $Y_1 + \dots + Y_k = n_0$, we have sampled $\mathbf{D}(n_0, k)$,

the distribution of component counts induced from taking all $p(n_0, k)$ possible assemblies equally likely. Note that when using (25), it is possible that a single Y value will be larger than n ; if one were to be horrified by such an occurrence, and tempted to throw away such samples, it would be preferable to use (24) instead.

Finally, if the goal is to sample *exactly* from the distribution of $\mathbf{D}(n, k)$, one strategy is “hard rejection/acceptance” sampling, which in this context would repeatedly proposing a value of (Y_1, \dots, Y_k) , testing to see if $Y_1 + \dots + Y_k = n$, and accepting if so, otherwise restarting. With $p(x, k, n) := \mathbb{P}_x(Y_1 + \dots + Y_k = n)$, the expected number of proposals before finding an acceptable one is $1/p(x, k, n)$, so the extended saddle heuristic, which corresponds to picking x to maximize $p(x, k, n)$ so that the *unconditioned* k -sample closely resembles the conditioned distribution, also serves as a recipe for relatively efficient simulation. However, it is possible to do much better than taking $1/p(x, k, n)$ independent proposals per achieved sample from the exact conditioned distribution, using probabilistic-divide-and-conquer, see [7, 20].

We should mention an alternate strategy, which is not limited to assemblies. Namely, the two-variable generating function in (18) corresponds to a two-parameter family of distributions for independent Z_1, Z_2, \dots , see [9, Section 8]. Here we take mixed notation: dummy variable z in the generating function corresponds to x when tilting, i.e., biasing with respect to x^{T_n} , (the proof of Lemma 3.2 is an example of the use of this), but we use θ for *both* roles, first as the dummy variable in the generating function in (18), *and* as second the tilting parameter, when biasing with respect to θ^{K_n} , with $T_n := Z_1 + 2Z_2 + \dots + nZ_n$ and $K_n := Z_1 + Z_2 + \dots + Z_n$. These tilted distributions have the property, similar to (22), that for every (θ, x) ,

(30)

$$\mathbf{CR}: \quad \mathcal{L}((D_1(n, k), D_2(n, k), \dots)) = \mathcal{L}_{\theta, x}((Z_1, Z_2, \dots) \mid T_n = n, K_n = k).$$

Relation (30) applies not just to assemblies, but also to multisets and selections. For *selections*, as characterized by (4), the (θ, x) distribution for Z_i is $\text{Binomial}(m_i, \theta x^i / (1 + \theta x^i))$. For *multisets*, as characterized by (5), the (θ, x) distribution for Z_i is $\text{Negative Binomial}(m_i, \theta x^i)$, so that the distribution of Z_i is the m_i -fold convolution of the Geometric (starting from zero, with ratio θx^i) distribution on $\{0, 1, 2, \dots\}$, that is, the distribution of G with $\mathbb{P}(G \geq k) = (\theta x^i)^k, k = 0, 1, 2, \dots$. For *assemblies*, the (θ, x) distribution for Z_i is Poisson with parameter $\lambda_i(\theta, x) = \theta m_i x^i / i!$; this is an extension of (14). Notice that for use in (23), the Poisson process construction for guaranteeing k arrivals, substituting $\lambda_i(\theta, x)$ for $\lambda_i(x)$ yields no change, as the factor θ *cancels* from numerator and denominator — in the second case of (23), the fraction naturally changes to $\lambda_i(\theta, x) / \sum_{j \geq 1} \lambda_j(\theta, x) = (\theta m_i x^i / i!) / (\theta M(x))$.

The extended saddle point heuristic says that, with parameters θ, x chosen so that $\mathbb{E}_{\theta,x} K_n$ is near k and $\mathbb{E}_{\theta,x} T_n$ is near n , then the conditioning on the right hand side of (22) has a mild effect, so the unconditioned (Z_1, Z_2, \dots, Z_n) is, with respect to various functionals, a good approximation to the distribution of $\mathbf{D}(n, k)$. Hence, without asking for theorems to certify how good an approximation one gets, there is the following:

The k -Boltzmann sampler, version 2. To quickly generate random surrogates for the component structure $\mathbf{D}(n, k)$, find parameters θ, x so that $\mathbb{E}_{\theta,x} K_n$ is near k and $\mathbb{E}_{\theta,x} T_n$ is near n , generate the independent process (Z_1, Z_2, \dots, Z_n) under its (θ, x) law, and accept these counts, even if $Z_1 + 2Z_2 + \dots + nZ_n$ is not equal to n , or $Z_1 + Z_2 + \dots + Z_n$ is not equal to k .

3. APPLICATION: LOW RANK STRUCTURES

In the context of the Tutte polynomial and its rank and nullity expansion, the *rank* of a graph with n vertices and k connected components is defined to be $r := n - k$. Graphs are an instance of a decomposable combinatorial structure, and since we are working with situations such as $k = n - \lfloor t\sqrt{n} \rfloor$, in which it is convenient to focus on $r = n - k$ and simply write $r \sim t\sqrt{n}$, we shall henceforth use the term *rank* for $n - k$ in the broader context of decomposable structures of size n with k components.

The extreme case of a low rank structure is the case $r = 0$, which forces all components to have size 1. Rank $r = 1$ forces the structure to have a single component of size 2, and all other components of size 1. Jumping up a little, rank $r = 3$ implies that the component structure be one of three possible types, namely $\lambda = (2, 2, 2, 1, 1, \dots, 1)$ or $\lambda = (3, 2, 1, 1, \dots, 1)$ or $\lambda = (4, 1, 1, \dots, 1)$. Clearly, one imagines removing 1 from each part. So we define the *copartition* of a partition λ to be the partition $\bar{\lambda}$ formed from λ by removing parts of size 1 and reducing each remaining part by 1 — that is, erasing the first column (or row, depending on one's choice of orientation) from the Ferrers diagram. For future reference,

$$(31) \quad \lambda = (\lambda_1, \lambda_2, \dots, \lambda_k) \text{ with } \lambda_j > 1, \lambda_{j+1} = \dots = \lambda_k = 1$$

has copartition

$$(32) \quad \bar{\lambda} = (\lambda_1 - 1, \lambda_2 - 1, \dots, \lambda_j - 1).$$

Thus, the three examples we gave for rank 3 correspond to $\bar{\lambda} = (1, 1, 1)$ and $\bar{\lambda} = (2, 1)$ and $\bar{\lambda} = (3)$. In general, for any $0 \leq r < n$, the number of component types for a structure of size n with rank r is p_r , the number of integer partitions of r , whose asymptotics as $r \rightarrow \infty$ are given by the Hardy–Ramanujan formula (6).

Partitions of n , with rank r , are in one to one correspondence with partitions of r ; the largest component of λ is one more than the largest component of $\bar{\lambda}$. Hence, as $r, n \rightarrow \infty$ with $n > r$ but otherwise logically independent of the relative growth rates of r and n , the size of the largest component of a

rank r partition of the integer n grows like (one plus) the largest part of a random integer partition of r . Erdős and Lehner [23] described the growth of the largest part of a random n partition:

$$(33) \quad \text{with } c = \pi/\sqrt{6}, \text{ as } n \rightarrow \infty, L_1 \sim \frac{1}{2c}\sqrt{n} \log(n)$$

in probability, so sampling uniformly from the $p(n, k) = p(n, n - r)$ integer partitions of rank r we have

$$(34) \quad \text{as } n, r \rightarrow \infty, L_1 \sim \frac{1}{2c}\sqrt{r} \log(r).$$

This behavior is in sharp contrast with the universal behavior for low rank assemblies, including set partitions, given by Theorems 3.3 – 3.14.

Open Problem 3.1. *Consider polynomials over \mathbb{F}_q , as in (7), in the low rank regime: pick uniformly from the $p(n, k)$ monic polynomials of degree n having exactly $k = n - r$ irreducible factors, where $r \sim \sqrt{n}$, or more generally $r \sim tn^\alpha$ for $\alpha \in (0, 1)$ with fixed $t \in (0, \infty)$. Determine the behavior of L_1 , the largest degree of an irreducible factor.*

For Question 3.1 above, it is easy to see that the behavior described in Theorem 3.3 does not hold. This is striking, since random polynomials over \mathbb{F}_q , and random permutations, behave very similarly with respect to typical aspects of the component structure. The key difference is that, with respect to component structure, permutations are an *assembly*, while polynomials are a *multiset*.

3.1. Low rank assemblies. Recall from Section 2.1 that the component structure of *assemblies* of size n with k components can be handled using independent random \mathbb{N} -valued variables Y_1, \dots, Y_k , conditional on $Y_1 + \dots + Y_k = n$. Since we focus on *rank* and the associated copartition, we consider $X_j := Y_j - 1$, so that as events,

$$(35) \quad \{Y_1 + \dots + Y_k = n\} = \{X_1 + \dots + X_k = r\}.$$

Assume $m_1 > 0$ and M has a nonzero radius of convergence. It is then immediate that as $x \rightarrow 0$, $M(x) \sim m_1 x$.

For any positive x less than the radius of convergence of $M(\cdot)$, the x -distribution of $X := Y - 1$ is given by

$$(36) \quad p_i \equiv p_i(x) \equiv \mathbb{P}_x(X = i) = \frac{m_1 x}{M(x)} \frac{m_{i+1} x^i}{m_1 (i+1)!}, \quad i = 0, 1, 2, \dots$$

In (36), the choice to also factor out m_1 is so that we may write

$$(p_0, p_1, p_2, \dots) \propto (1, \frac{m_2}{2m_1}x, \frac{m_3}{6m_1}x^2, \dots),$$

with the constant of proportionality being $\frac{m_1 x}{M(x)} \rightarrow 1$ as $x \rightarrow 0+$.

Lemma 3.2. *Assume that $m_1 > 0$ and M has radius of convergence $R \in (0, \infty]$. As $x \rightarrow 0+$,*

$$(37) \quad \mathbb{E}_x(X) \sim \frac{m_2}{2m_1} x,$$

and also

$$\mathbb{P}(X = 0) \rightarrow 1, \quad \mathbb{P}(X = 1) \sim \mathbb{E}_x X, \quad \text{and} \quad \mathbb{P}(X \geq 2) = O(x^2).$$

Proof. Starting from (10), let $A(z) := M(z)/z = \sum_{i \geq 0} m_{i+1} z^i / (i+1)! = m_1 + (m_2/2)z + (m_3/6)z^2 + \dots$. The probability generating function for the x -distribution of X in (36), expressed with dummy variable z , is

$$G(z) \equiv G_{x,X}(z) := \mathbb{E}_x z^X = \frac{A(xz)}{A(x)}$$

with $G'(z) = xA'(xz)/A(x)$, hence $\mathbb{E}_x X = G'(1) = xA'(x)/A(x)$. Obviously, as $x \rightarrow 0$, $A(z) \rightarrow m_1$ and $A'(z) \rightarrow m_2/2$. This establishes (37), and the remaining claims are obvious. \square

Now take X, X_1, \dots, X_k to be i.i.d., and for $i \geq 0$, let

$$(38) \quad N_i := \sum_{j=1}^k 1(X_j = i)$$

be the number of i s in the sample of size k . Note that for all outcomes, $N_0 + N_1 + \dots = k$. *Warning:* for this application to low rank only, we shift notation used in (25), without changing the letter N , from $N_i \equiv N_i(k) :=$ the count, how many of Y_1, \dots, Y_k are equal to i , for $i \geq 1$, to $N_i :=$ the count, how many of X_1, \dots, X_k are equal to i , for $i \geq 0$. Hence the conditioning event (35) will now be expressed as

$$(39) \quad \{\omega : X_1 + \dots + X_k = r\} = \{\omega : 0N_0 + N_1 + 2N_2 + \dots = r\}.$$

The joint distribution of (N_0, N_1, \dots) is similar to a multinomial distribution, and indeed, for any list of m disjoint sets B_1, \dots, B_m whose union is \mathbb{Z}_+ , the *lumped together* count vector $(\overline{N}_1, \dots, \overline{N}_m)$ where $\overline{N}_j := \sum_{i \in B_j} N_i$, has a genuine multinomial distribution, corresponding to k tosses of an m -sided die, in which face j has probability $\overline{p}_j := \sum_{i \in B_j} p_i$. A particular case of the preceeding remark, with $m = 2$ and $B_1 =$ the singleton $\{i\}$, is that the distribution of N_i is $\text{Binomial}(k, p_i)$.

Our goal is, for an appropriate choice of x , to first approximate $\mathbb{P}_x(N_1 + 2N_2 + \dots = r)$ and then approximate the conditional distribution of (N_0, N_1, N_2, \dots) given that $N_1 + 2N_2 + \dots = r$ by its unconditional distribution.

3.2. The critical regime for having components of size 3. We have already assumed that $m_1 > 0$ and M has a strictly positive radius of convergence, and now we *also* assume that $m_2, m_3 > 0$.

Fix $t \in (0, \infty)$ and fix any sequence $k(1), k(2), \dots$ such that $r \equiv r(n) := n - k$ satisfies

$$(40) \quad r \sim t\sqrt{n}.$$

Observe that this entails $r = o(n)$ and hence $k \sim n$, so we also have $r \sim t\sqrt{k}$. In particular, we have $r/k \rightarrow 0$ as $n \rightarrow \infty$. We want to find x so that (36) has $kp_1 = r$. Lemma 3.2 implies that for small x , $p_1 \sim \mathbb{E}_x X \sim m_2 x / (2m_1)$, so the first guess $x_0 = (2m_1/m_2)r/k$ would have $kp_1(x_0) \sim r$. We have $x_0 \rightarrow 0$, so for sufficiently large n , $x_0 < R/2$, and using $m_1 x / M(x) \rightarrow 1$, we can find x relatively close to x_0 , such that $kp_1(x) = r$.

For sufficiently large n , define

$$(41) \quad x \equiv x(n) = \text{the solution of } p_1 = \frac{r}{k}, \text{ so that } \mathbb{E}_x N_1 = kp_1 = r,$$

and in case there is more than one positive solution, amend (41) to choose the smallest positive solution. It follows, from Lemma 3.2 combined with (40), that this $x \equiv x(n)$ satisfies

$$x \sim \frac{2m_1}{m_2} \frac{r}{k} \sim \frac{2m_1}{m_2} \frac{t}{\sqrt{k}}$$

and hence

$$(42) \quad kp_2 \sim k \frac{m_3}{6m_1} x^2 \rightarrow \frac{2m_1 m_3}{3m_2^2} t^2.$$

We now sketch the overall argument for Theorem 3.3, saving minor details for the formal proof. From (42) we see that N_2 , whose marginal distribution is exactly $\text{Binomial}(k, p_2)$, will be approximately Poisson with mean given by the r.h.s. of (42). We have $\mathbb{E}(3N_3 + 4N_4 + \dots) = O(x)$ times $\mathbb{E}N_2$, so those contributions will be negligible. Thus, in typical outcomes, with *small* exceptional probability, both $K_0 := N_2 + N_3 + \dots$ and $R_0 := 2N_2 + 3N_3 + \dots$ will be zero, or small positive integers. Conditional on $K_0 = k_0$, and also conditioning on which indices in the k -sample contributed to K_0 , there are $k - k_0$ rolls of the die, on which the event in (39) occurs if and only if each of those $k - k_0$ rolls shows face 0 or 1, and the total number of 1s is $r - r_0$. But this is just asking for $\text{Binomial}(k - k_0, p_1/(p_0 + p_1))$ to have some value $r - r_0$ close to its mean, and the conditional probability is close to $1/\sqrt{2\pi k p_1}$, regardless of which small values k_0, r_0 . So the overall probability of the event in (39) is asymptotic to $1/\sqrt{2\pi k p_1}$, and conditional on that event, N_2 is still close to Poisson with mean given by the right side of (42). This shows that in the earlier consideration, the *small* probability for the exceptional event needs to be $o(1/\sqrt{r})$, and this is sufficient to prove (44). Recall that via the shift $X = Y - 1$, the count N_2 here actually corresponds to blocks of size 3.

Theorem 3.3. *Consider an assembly as governed by (10) and (13), and assume further that $m_1, m_2, m_3 > 0$ and $M(\cdot)$ has a strictly positive radius of convergence. Fix $t > 0$ and a sequence $k(1), k(2), \dots \geq 1$ such that $n - k(n) \sim t\sqrt{n}$. Given n , pick an assembly uniformly from the $p(n, k)$*

choices having exactly k components. Then, with probability tending to 1, the largest component L_1 has size 2 or 3, the number of components of size 3 has distributional limit given by

$$(43) \quad D_3(n, k) \rightarrow^d \text{Poisson with mean } \lambda \equiv \lambda(t, M) := \frac{2m_1m_3}{3m_2^2}t^2,$$

and hence for an assembly of size n chosen uniformly from the $p(n, k)$ possibilities with k components, we have

$$(44) \quad \mathbb{P}(L_1 = 2) \rightarrow \exp(-\lambda(t, M)).$$

The error in the approximation (44), and indeed the total variation distance between the ingredients in (43), i.e., $d_{\text{TV}}(D_3(n, k), \text{Poisson}(\lambda))$, is at most $O_t((\log^2 n)/\sqrt{n})$.

Proof. Pick x as per (41) modified slightly, so that

$$k \frac{p_1(x)}{p_0(x) + p_1(x)} = r.$$

The new choice of x is asymptotic to the old choice, and we still have a result like (42), namely that $\mathbb{E}_x N_2 = O_t(1)$. The implicit constant in the big O depends both on t and $M(\cdot)$, but we highlight the only the dependence on t , since we consider the assembly M as fixed. The binomial distribution of N_2 satisfies the Hoeffding bound (see [4, equation (41)]), for all $y > 1$, $\mathbb{P}(N_2 \geq y \mathbb{E} N_2) \leq y^{-y \mathbb{E} N_2} e^{(y-1) \mathbb{E} N_2}$, so we can pick $r_2 = O_t(\log n)$ and $n_2 \equiv n_2(t)$ so that the *bad* event $B_2 := \{2N_2 \geq r_2\}$ has $\mathbb{P}_x(B_2) < 1/n$ for all $n > n_2$. Similarly, but not as delicate, one can pick $r_3 = O_t(\log n)$ and $n_3 \equiv n_3(t)$ so that the *bad* event $B_3 := \{3N_3 + 4N_4 + \dots \geq r_3\}$ has $\mathbb{P}_x(B_3) < 1/n$ for all $n > n_3$. Combining, the bad event $B := B_2 \cup B_3$ has $\mathbb{P}_x(B) < 2/n$ for all sufficiently large n , and on the complementary event, B^c , with $k_1 = r_1 := r_2 + r_3 = O_t(\log n)$ we have $K_0 := N_2 + N_3 + \dots \leq k_1$ and $R_0 := 2N_2 + 3N_3 + \dots \leq r_1$.

Conditional on $\{K_0 = k_0\}$ and further conditioning on *which* of the $k - k_0$ indices $j \in [k]$ did not contribute to K_0 , i.e., those j for which $X_j = 0$ or 1, we have $k - k_0$ independent trials where $p := \mathbb{P}(X_j = 1) = 1 - \mathbb{P}(X_j = 0) = p_1/(p_0 + p_1)$, and from (41) we have $p \sim t/\sqrt{k}$. For the binomial(k, p) distribution, the target r is exactly the mean, the binomial point probability at the mean is approximately $1/\sqrt{2\pi k p} = 1/\sqrt{2\pi r}$, with error controlled by Stirling's formula; the relative error is $O(1/\sqrt{k})$. Likewise, for the perturbations, where k is replaced by $k - k_0$ and the target is replaced by $r - r_0$, with $0 \leq k_0, r_0 \leq k_1 = O_t(\log k)$, (recalling that $k \sim n$), the probability that Binomial($k - k_0, p$) hits the point $r - r_0$ is asymptotically $1/\sqrt{2\pi r}$. For the relative error, the main contribution comes from the target being at most r_0 from the center, and since the variance is order of r , the resulting relative error is $O(r_0^2/r) = O_t((\log n)^2/\sqrt{n})$.

The second paragraph of this proof shows that the *contribution* to $\mathbb{P}_x(N_1 + 2N_2 + \dots = r)$ from the event B^c is asymptotically $1/\sqrt{2\pi r}$, with relative

error at most $O_t((\log^2 n)/\sqrt{n})$. Our bound $\mathbb{P}_x(B) < 2/n$ from the first paragraph of this proof is of a smaller order, so the net result is that $\mathbb{P}_x(N_1 + 2N_2 + \dots = r) \sim 1/\sqrt{2\pi r}(1 + O_t((\log^2 n)/\sqrt{n}))$. Now that $\mathbb{P}_x(N_1 + 2N_2 + \dots = r)$ has been estimated asymptotically, the same argument from the second paragraph shows that the *conditional* probability that $N_2 = m$ given that $N_1 + 2N_2 + \dots = r$ is relatively close to the unconditional probability, again with relative error that is $O_t((\log^2 n)/\sqrt{n})$. Finally, the marginal distribution of N_2 is exactly $\text{Binomial}(k, p_1(x))$, with mean asymptotically $O_t(1)$, as given in detail by (42) — so the total variation distance from this binomial marginal distribution, to its Poisson approximation, is $O(1/k) = O_t(1/n)$. This proves both (43) and (44) and even shows that $\mathbb{P}(L_1 = 2) - \exp(-\lambda(M, t)) = O_t((\log^2 n)/\sqrt{n})$. \square

Corollary 3.4. *Consider an assembly as governed by (10) and (13), and assume further that $m_1, m_2, m_3 > 0$ and $M(\cdot)$ has a strictly positive radius of convergence. Fix $t > 0$ and a sequence $k(1), k(2), \dots \geq 1$ such that $r := n - k(n) \sim t\sqrt{n}$. Then*

$$(45) \quad p(n, k) \sim \frac{n^{2r} m_1^{n-2r} m_2^r}{r! 2^r} \exp\left(-t^2 \left(2 - \frac{2m_1 m_3}{3m_2^2}\right)\right),$$

and the relative error in (45) is $O_t((\log^2 n)/\sqrt{n})$.

Proof. With rank $r := n - k$, let $\mathbf{a} = (n - 2r, r, 0, 0, \dots)$. Note, this specifies the integer partition with $(n - 2r) + r = k$ parts, and is a partition of $1 \times (n - 2r) + 2 \times r = n$, so $N(n, \mathbf{a})$ is one of the contributions to $p(n, k)$, as discussed in the first paragraph of Section 2. From (11), and writing the falling power x falling i as $(x)_i$, we have

$$(46) \quad N(n, \mathbf{a}) = \frac{n! m_1^{n-2r} m_2^r}{(n - 2r)! r! 2^r} = \frac{(n)_{2r} m_1^{n-2r} m_2^r}{r! 2^r}.$$

Using $r \sim t\sqrt{n}$ and the usual asymptotic for the birthday problem, that $(n)_i/n^i \sim \exp(-i^2/(2n))$ whenever $i = o(n^{2/3})$, we have $(n)_{2r}/n^{2r} \rightarrow \exp(-2t^2)$. Notice that, on the left side of (44), we have

$$\mathbb{P}(L_1 = 2) = \frac{N(n, \mathbf{a})}{p(n, k)}, \text{ so that } p(n, k) = \frac{N(n, \mathbf{a})}{\mathbb{P}(L_1 = 2)}.$$

Combining this with (43) and (44) yields the desired result. \square

Remark 3.5. *The upper bound on the relative error, proved in Theorem 3.3 and inherited by Corollary 3.4, is $O_t((\log^2 n)/\sqrt{n})$. This reflects our desire to be succinct. We believe that the true error is order $1/\sqrt{n}$, and will make formal conjectures out of this, with Conjectures 3.6 — 3.8. Note that we are working under the regime $r \sim t\sqrt{n}$, so $1/\sqrt{n} \sim t/r$.*

Conjecture 3.6. *Under the hypotheses of Theorem 3.3, the result (44) can be improved to*

$$(47) \quad \mathbb{P}(L_1 = 2) = \exp(-\lambda(t, M)) (1 + O_t(1/r))$$

$$= \exp(-\lambda(t, M)) (1 + O_t(1/\sqrt{n})).$$

Conjecture 3.7. *Under the hypotheses of Theorem 3.3, the true order of error in (47) is order of $1/r$, in the sense that there is a function $C : (0, \infty) \rightarrow \mathbb{R}$, depending on m_1, m_2, m_3 , such that when $n \rightarrow \infty$ and $r = \lfloor t\sqrt{n} \rfloor$, we have*

$$(48) \quad \mathbb{P}(L_1 = 2) = \exp(-\lambda(t, M)) + \frac{C(t)}{r} + o_t(1/r).$$

Conjecture 3.8. *The function $C(\cdot)$ for use in (48) is given explicitly by*

$$(49) \quad C(t) = (2\lambda^2 + \lambda) - 2t^2\lambda(t, M) - t^2 \frac{m_4}{4} \lambda.$$

Remark 3.9. *The expression in (49), albeit highly technical, is a plausible attempt to name all the order of $1/r$ contributions to the relative error between $\mathbb{P}(L_1 = 2)$ and $\exp(-\lambda(t, M))$. We view $C(t)$ as a sum with three terms.*

For the first term, $(2\lambda^2 + \lambda)$, consider outcomes where $L_1 \leq 3$; these are $(N_2 = j, N_1 = r - 2j, N_0 = k - r - j)$, for $j = 0, 1, 2, \dots$. In the second paragraph of the proof of Theorem 3.3, these correspond to the situation where $k_0 = j$, the conditional distribution of N_1 is exactly $\text{Binomial}(k - j, r/k)$, and the target value is $r - 2j$. The relative error between $\text{Binomial}(k - j, r/k)[r - 2j]$ and $\text{Binomial}(k, r/k)[r - 2j]$ is $O(1/k) = o(1/r)$, negligible here. For $\text{Binomial}(k, r/k)$, the relative difference between the mass at r and the mass at $r - m$, i.e., between $\text{Binomial}(k, r/k)[r]$ and $\text{Binomial}(k, r/k)[r - m]$, is $1 - (m)_r/m^r + o(1/r) = \binom{m-1}{2}/r + o(1/r)$. We use this with $m = 2j = 2N_2$. Under the Poisson approximation where the distribution of N_2 is taken to be $\text{Poisson}(\lambda)$, we have $\mathbb{E} \binom{2N_2-1}{2} = 2\lambda^2 + \lambda$.

For the second term, consider that the choice used for x in the proof of Theorem 3.3, which is described even more explicitly by (55), leads to $\lambda' := \mathbb{E} N_2 = n/(n - 2r) \times \lambda$. The relative error in approximating $\exp(-\lambda')$ by $\exp(-\lambda)$ is $2t^2\lambda/r + o(1/r)$.

For the third term, consider the event $N_3 = 1$, corresponding to the assembly having exactly one component of size 4. In Remark 3.10, the most likely representative of this event is described by \mathbf{a}''' , leading to the plausible belief that, with the event in (39) denoted as G ,

$$\frac{\mathbb{P}(N_3 = 1, G)}{\mathbb{P}(G)} \sim \frac{N(n, \mathbf{a}''')}{N(n, \mathbf{a})} \sim \frac{m_4}{4} \frac{r}{n} \sim t^2 \frac{m_4}{4} \lambda \frac{1}{r}.$$

Remark 3.10. *To give perspective on the meaning and extent of sharpness of the upcoming Theorem 3.13, we consider four particular partition types for an assembly of size n to have rank r . In each case, we describe the rank r partition of n first by its counts \mathbf{a} , then via the notation $\lambda = 1^{a_1} 2^{a_2} \dots$, and finally by the copartition $\bar{\lambda}$ as described by (31) and (32). The first type*

is familiar from the proof of Corollary 3.4.

$$\begin{array}{llll}
\mathbf{a} &= (n-2r, r, 0, \dots) & \lambda &= 1^{n-2r} 2^r & \bar{\lambda} &= 1^r \\
\mathbf{a}' &= (n-2r+1, r-2, 1, 0, \dots) & \lambda' &= 1^{n-2r+1} 2^{r-2} 3^1 & \bar{\lambda}' &= 1^{r-2} 2^1 \\
\mathbf{a}'' &= (n-2r+2, r-4, 2, 0, \dots) & \lambda'' &= 1^{n-2r+2} 2^{r-4} 3^2 & \bar{\lambda}'' &= 1^{r-4} 2^2 \\
\mathbf{a}''' &= (n-2r+2, r-3, 0, 1, 0, \dots) & \lambda''' &= 1^{n-2r+2} 2^{r-3} 4^1 & \bar{\lambda}''' &= 1^{r-3} 3^1
\end{array}$$

The exact count of how many M -assemblies have type \mathbf{a} , the first case in the list above, is given in (46). The corresponding exact counts for the next three cases are

$$(50) \quad N(n, \mathbf{a}') = \frac{(n)_{2r-1} m_1^{n-2r+1} m_2^{r-2} m_3}{(r-2)! 2^{r-2} 3!}.$$

$$(51) \quad N(n, \mathbf{a}'') = \frac{(n)_{2r-2} m_1^{n-2r+2} m_2^{r-4} m_3^2}{(r-4)! 2^{r-4} 2! (3!)^2}.$$

$$(52) \quad N(n, \mathbf{a}''') = \frac{(n)_{2r-2} m_1^{n-2r+2} m_2^{r-3} m_4}{(r-3)! 2^{r-3} 4!}.$$

Considering the ratios of each of the above three with $N(n, \mathbf{a})$, which for $\mathbf{a} = (n-2r, r, 0, \dots)$ is the exact count of M -assemblies of rank r and with $L_1 = 2$, for any $r \geq 1$, we see that

$$\frac{N(n, \mathbf{a}')}{N(n, \mathbf{a})} = \frac{m_1 (r)_2 2^2 m_3}{(n-2r+1) m_2^2 3!},$$

$$\frac{N(n, \mathbf{a}'')}{N(n, \mathbf{a})} = \frac{m_1^2 (r)_4 2^4 m_3^2}{(n-2r+1)(n-2r+2)(3!)^2 m_2^4},$$

$$\frac{N(n, \mathbf{a}''')}{N(n, \mathbf{a})} = \frac{m_1^2 (r)_3 2^3 m_4}{(n-2r+1)(n-2r+2) m_2^3 4!}.$$

In particular, if $r \rightarrow \infty$ and $r = o(n)$ then, with the symbol \asymp used to mean that the ratio is bounded away from zero and infinity, we have

$$(53) \quad \frac{N(n, \mathbf{a}')}{N(n, \mathbf{a})} \sim \frac{2m_1 m_3}{3m_2^2} \frac{r^2}{n} \asymp \frac{r^2}{n}, \quad \frac{N(n, \mathbf{a}'')}{N(n, \mathbf{a})} \asymp \frac{r^4}{n^2}, \quad \frac{N(n, \mathbf{a}''')}{N(n, \mathbf{a})} \asymp \frac{r^3}{n^2}.$$

This will show that the error bound in Theorem 3.13 is sharp. It hints at the job of Lemma 3.11, which is to compare $N(n, \mathbf{a})$ with the combined count of all rank r assemblies of size n having only parts of size 1, 2, and 3, by direct combinatorial argument. And it gives perspective to Lemma 3.12, which uses the saddle point approximation to give an upper bound on all cases, like \mathbf{a}''' , involving at least one part of size 4 or greater.

Lemma 3.11. *Consider M -assemblies with $m_1, m_2 > 0$. Assume $0 < r < n/2$ and let*

$$(54) \quad y = \frac{2m_1 m_3}{3m_2^2} \frac{r^2}{n - 2r}.$$

Picking uniformly from the $p(n, n - r)$ M -assemblies of rank r , we have

$$\mathbb{P}(L_1 = 3) \leq e^y - 1.$$

Proof. Let

$$\mathbf{a}^{(j)} := (n - 2r + j, r - 2j, j, 0, 0, \dots, 0),$$

so that the rank r partitions \mathbf{a}, \mathbf{a}' , and \mathbf{a}'' in Remark 3.10 are exactly $\mathbf{a}^{(j)}$ for $j = 0, 1, 2$. Since $r > 0$, we cannot have $L_1 = 1$, and the event $(L_1 \leq 3)$ is precisely the event $(\mathbf{D}(n, n - r) = \mathbf{a}^{(j)})$ for some j with $0 \leq j \leq r/2$. As in the proof of Corollary 3.4, the event $(L_1 = 2)$ is precisely the event $(\mathbf{D}(n, n - r) = \mathbf{a}^{(0)})$.

From (11) we have, for $0 \leq j \leq r/2$,

$$N(\mathbf{a}^{(j)}, n) = n! \frac{m_1^{n-(2r-j)} m_2^{r-2j} m_3^j}{(n - (2r - j))! 2^{r-2j} (r - 2j)! (3!)^j j!}$$

so that

$$\frac{N(\mathbf{a}^{(j)}, n)}{N(\mathbf{a}^{(0)}, n)} = \frac{m_1^j (r)_{2j} 2^{2j} m_3^j}{(n - 2r + j)_j (3!)^j m_2^{2j} j!} \leq y^j / j!,$$

hence

$$\mathbb{P}(L_1 = 3) \leq \sum_{1 \leq j \leq r/2} y^j / j! \leq e^y - 1.$$

□

In the next lemma, our goal is to give a completely effective lower bound on the probability of the event in (35), in a way that gives an asymptotically useful bound for the situation with $r = o(\sqrt{n})$. Our saddle choice for the value of the parameter x for use in (36) is determined by the requirement

$$k \frac{p_1(x)}{p_0(x) + p_1(x)} = r, \text{ equivalently } \frac{x}{2m_1/m_2 + x} = \frac{r}{n - r},$$

equivalently

$$(55) \quad x = \frac{2m_1 r}{m_2(n - 2r)}.$$

Observe that in any low rank regime, that is, whenever $r = o(n)$, we have $x \sim (2m_1/m_2) r/n$. We take

$$(56) \quad \rho := \sup_{i \geq 3} \left(\frac{m_i}{i!} \right)^{1/i},$$

noting that the assumption that M has strictly positive radius of convergence is equivalent to the condition that $\rho < \infty$. Observe that when $r = o(n)$ the left side of (58) is $2k\rho^3 x^2/m_1 \sim (2\rho^3/m_1) nx^2 \sim (2\rho^3/m_1) r^2/n$, so the

condition that $r = o(\sqrt{n})$ is sufficient to guarantee that (58) holds *eventually*. Finally, observe that for the situation of interest, which is $r = o(\sqrt{n})$, the r.h.s. of (60) is order of $kx^3\sqrt{r} \sim n(r/n)^3\sqrt{r} = r^{-1/2}(r^2/n)^2$, so compared with the error contribution from Lemma 3.11, the error contribution from Lemma 3.12 is of smaller order.

Lemma 3.12. *Consider an assembly as governed by (10) and (13), and assume further that $m_1, m_2 > 0$ and $M(\cdot)$ has a strictly positive radius of convergence. Given $n, r = n - k$ let the parameter in (36) be given by (55), and let ρ be given by (56). Assume that n, r satisfy*

$$(57) \quad x\rho \leq 1/2.$$

and

$$(58) \quad \frac{2k\rho^3x^2}{m_1} \leq 1/2.$$

Then, with $c_0 := e/\sqrt{2\pi}$,

$$\mathbb{P}(X_1 + \dots + X_k = r) \geq \frac{1}{2c_0\sqrt{2\pi r}}$$

and

$$(59) \quad \mathbb{P}(N_3 + N_4 + \dots > 0) \leq k\mathbb{P}(X \geq 3) \leq k\frac{2\rho^4x^3}{m_1}$$

and hence

$$(60) \quad \mathbb{P}(L_1 \geq 4) \leq k\frac{2\rho^4x^3}{m_1} 2c_0\sqrt{2\pi r} =: u_4(n, r).$$

Proof. Using (56) and the assumption that $x\rho \leq 1/2$, we have

$$\sum_{i \geq 3} \frac{m_i x^i}{i!} \leq \sum_{i \geq 3} (\rho x)^i = \frac{(\rho x)^3}{1 - \rho x} \leq 2\rho^3 x^3.$$

Hence in (36), with this choice of x , we have

$$p_0(x) + p_1(x) = \frac{m_1 x + m_2 x^2/2}{M(x)} > \frac{m_1 x}{m_1 x + 2\rho^3 x^3} = \frac{m_1}{m_1 + 2\rho^3 x^2}$$

so that

$$(61) \quad \mathbb{P}(X \geq 2) = 1 - (p_0 + p_1) \leq \frac{2\rho^3 x^2}{m_1 + 2\rho^3 x^2} \leq \frac{2\rho^3 x^2}{m_1}.$$

This yields

$$(62) \quad (p_0 + p_1)^k = (1 - \mathbb{P}(X \geq 2))^k \geq 1 - k\mathbb{P}(X \geq 2) \geq 1 - \frac{2k\rho^3 x^2}{m_1}.$$

With counts N_i , for $i \geq 0$, as specified just before (39), and recalling that $k = n - r$, and writing $p := p_1/(p_0 + p_1)$ so that by (55) we have also $p = r/k$, we have

$$\mathbb{P}(N_0 = n - 2r, N_1 = r) = \frac{k!}{(k - r)!r!} p_0^{k-r} p_1^r = \frac{k!}{(k - r)!r!} (1 - p)^{k-r} p^r (p_0 + p_1)^k.$$

The first factor on the r.h.s. above is a point probability for a binomial distribution with mean r , asymptotically $1/\sqrt{2\pi r(1-p)}$, and always at least $1/(c_0\sqrt{2\pi r})$, where $c_0 := e^1/\sqrt{2\pi} = 1.0844375514192\dots$. The second factor on the right side above is bounded via (62) and the hypothesis (58), hence

$$\mathbb{P}(X_1 + \dots + X_k = n) \geq \mathbb{P}(N_0 = n - 2r, N_1 = r) \geq \frac{1}{2c_0\sqrt{2\pi r}}.$$

Using (56) and the assumption that $x\rho \leq 1/2$, we have

$$\sum_{i \geq 4} \frac{m_i x^i}{i!} \leq \sum_{i \geq 4} (\rho x)^i = \frac{(\rho x)^4}{1 - \rho x} \leq 2\rho^4 x^4,$$

and $M(x) \geq m_1 x$, hence

$$\begin{aligned} \mathbb{P}(N_3 + N_4 + \dots > 0) &\leq k \mathbb{P}(X \geq 3) \\ &= \frac{k}{M(x)} \sum_{i \geq 4} \frac{m_i x^i}{i!} \leq \frac{2k\rho^4 x^4}{m_1 x}. \end{aligned}$$

To prove (60) we *overpower* the requirement for the occurrence of the conditioning event:

$$\begin{aligned} \mathbb{P}(L_1 \geq 4) &= \mathbb{P}(N_3 + N_4 + \dots > 0 | X_1 + \dots + X_k = r) \\ &= \frac{\mathbb{P}(N_3 + N_4 + \dots > 0 \text{ and } X_1 + \dots + X_k = r)}{\mathbb{P}(X_1 + \dots + X_k = r)} \\ &\leq \frac{\mathbb{P}(N_3 + N_4 + \dots > 0)}{\mathbb{P}(X_1 + \dots + X_k = r)} \\ &\leq k \frac{2\rho^4 x^3}{m_1} \frac{1}{2c_0\sqrt{2\pi r}}. \end{aligned}$$

□

Theorem 3.13. *Consider an assembly as governed by (10) and (13). Given n, r , with $k := n - r$ pick an assembly uniformly from the $p(n, k)$ choices having exactly k components.*

Assume the hypotheses of Lemmas 3.11 and 3.12, i.e., $m_1, m_2 > 0$, $M(\cdot)$ has a strictly positive radius of convergence, $0 < r < n/2$, and that x and ρ as given by (55) and (56) satisfy (57) and (58).

Then, with $y \equiv y(n, r)$ given by (54) and u_4 given by (60),

$$\mathbb{P}(L_1 \geq 3) \leq (e^y - 1) + u_4(n, r) =: z.$$

Hence, in case $z < 1$,

$$(63) \quad \frac{n^{2r} m_1^{n-2r} m_2^r}{r! 2^r} \leq p(n, k) \leq \frac{n^{2r} m_1^{n-2r} m_2^r}{r! 2^r} \left(1 + \frac{z}{1-z}\right).$$

Note that when $r = o(\sqrt{n})$, the upper bound on the relative error is

$$z/(1-z) \sim z \sim y \sim 2m_1 m_3 / (3m_2^2) r^2 / n \asymp r^2 / n.$$

Proof. The first statement is an immediate combination of the conclusions of Lemmas 3.11 and 3.12; the second statement follows by reasoning akin to that used in the proof of Corollary 3.4. The asymptotic analysis was given in the paragraph preceeding Lemma 3.12, and the calculation in (53) shows that the upper bound is asymptotically sharp. \square

Theorem 3.14. *Consider an assembly as governed by (10) and (13), and assume further that $m_1, m_2, \dots > 0$ and $M(\cdot)$ has a strictly positive radius of convergence. Fix a sequence $k(1), k(2), \dots \geq 1$ with $1 \leq k(n) \leq n$. Given n , pick an assembly uniformly from the $p(n, k)$ choices having exactly k components. Write $r = n - k$. Assume that for some $\varepsilon > 0$, $r = o(n^{1-\varepsilon})$.*

Then for $\ell = 1, 2, \dots$,

- *If $r = o(n^{\ell/(1+\ell)})$ then $\mathbb{P}(L_1 \leq \ell + 1) \rightarrow 1$.*
- *If $n^{\ell/(1+\ell)} = o(r)$ then $\mathbb{P}(L_1 > \ell + 1) \rightarrow 1$.*
- *If $\liminf \log_n r > \ell/(\ell + 1)$, then with x given by (41), for each i with $1 \leq i \leq \ell + 2$,*

$$(64) \quad 1 = \mathbb{P} \left(D_i(n, k) \sim k \frac{m_i x^{i-1}}{m_1 i!} \right).$$

- *If for fixed $t > 0$ we have $r \sim t n^{\ell/(1+\ell)}$ then, with*

$$(65) \quad \lambda \equiv \lambda(t, \ell, M) := \frac{2^{\ell+1} m_1^\ell m_{\ell+2}}{(\ell + 2)! m_2^{\ell+1}} t^{\ell+1},$$

$$(66) \quad \mathbb{P}(L_1 = \ell + 1) \rightarrow e^{-\lambda} \text{ and } \mathbb{P}(L_1 = \ell + 2) \rightarrow 1 - e^{-\lambda}.$$

Proof. We will sketch two computations that differ from the situation of Theorems 3.3 and 3.13. The remaining details for all claims are similar to arguments given in the proof of Theorem 3.3, although not as delicate, and we shall omit the details.

The key computation for the borderline behavior in (66) is that (41) entails

$$x \sim \frac{2m_1}{m_2} \frac{r}{k} \sim \frac{2m_1}{m_2} \frac{t}{k^{1/(1+\ell)}}$$

hence $\mathbb{E} N_{\ell+1}$ is asymptotic to

$$k p_{\ell+1} \sim k \frac{m_{\ell+2}}{m_1(\ell + 2)!} x^{\ell+1} \sim \frac{2^{\ell+1} m_1^\ell m_{\ell+2}}{(\ell + 2)! m_2^{\ell+1}} t^{\ell+1} =: \lambda(t, \ell, M).$$

Recall that $N_{\ell+1}$ counts how many of the X_1, \dots, X_k are equal to $\ell + 1$, which is the same as the number of Y_1, \dots, Y_k which are equal to $\ell + 2$.

For (64), consider a subsequence along which $\alpha = \lim \log_n r \in (\ell/(\ell + 1), 1)$. By (41) and Lemma 3.2, $x \sim \frac{2m_1}{m_2} \frac{r}{k} \approx n^{\alpha-1}$ (with the usual large deviation theory notation, $a_n \approx b_n$ to mean that $\log a_n \sim \log b_n$) and hence

$$\mathbb{E} N_{i-1} = k \frac{m_1 x}{M(x)} \frac{m_i x^{i-1}}{m_1 i!} \sim k \frac{m_i x^{i-1}}{m_1 i!}$$

so that $\mathbb{E} N_{i-1} \approx n^{1+(\alpha-1)(i-1)} = n^\delta$, with $\delta = 1 + (i-1)(\alpha-1) > 0$ using $i-1 \leq \ell+1$ and $1-\alpha > 1/(\ell+1)$. The marginal distribution of N_{i-1} is Binomial($k, p_{i-1}(x)$), and a moderate deviation bound, that the probability of being more than $c \log n$ standard deviations away from the mean is $o(1/n)$, in conjunction with an overall argument that $\mathbb{P}(X_1 + \dots + X_k = r) \asymp 1/\sqrt{r}$, establishes (64). (One could prove a stronger version of (64), allowing for example $r \sim n^{\ell/(\ell+1)} \log \log \log n$, but then for $i = \ell+2$, $\mathbb{E} N_{i-1}$ would grow very slowly, and instead of brute force “overpowering the conditioning”, one would have to argue some approximate independence between N_{i-1} and the event $X_1 + \dots + X_k = r$, as we did in the proof of Theorem 3.3.) \square

3.3. A completely effective version of Theorem 3.3. Theorem 3.3 establishes the asymptotic behavior in the critical regime for having components of size 3. In this section, we wish to highlight that by equally elementary but slightly more tedious calculations, one can just as easily provide quantitative bounds, i.e., completely effective inequalities for the relevant probabilities for all finite values of the parameters, which yield the asymptotic behavior as a corollary. We utilize the following lemmas in the proof of Theorem 3.19.

Lemma 3.15. [25, Section VI.10, Problem 34] *Suppose $0 < p < 1$ and $n \in \mathbb{N}$. Let $\lambda = np$, and define $b(k; n, p) := \binom{n}{k} p^k (1-p)^{n-k}$, and $p(k; \lambda) := \frac{\lambda^k}{k!} e^{-\lambda}$. Then we have*

$$p(k; \lambda) e^{-\frac{k^2}{n-k} - \frac{\lambda^2}{n-\lambda}} < b(k; n, p) < p(k; \lambda) e^{k\lambda/n}.$$

Lemma 3.16. *Suppose $0 < p < 1$, $pn \geq 1$ and $0 < k = pn + h < n$. Put*

$$\beta = \frac{1}{12k} + \frac{1}{12(n-k)},$$

Let $b(k; n, p)$ denote the point probability that a Binomial random variable with parameters n and p is equal to k , and $q = 1 - p$. We have

$$\sqrt{2\pi p q n} b(k; n, p) < \exp \left(-\frac{h}{2pn} + \frac{h}{2qn} - \frac{h^2}{pn} - \frac{h^2}{qn} \right).$$

We also have

$$\sqrt{2\pi p q n} b(k; n, p) > \begin{cases} \exp \left(-\beta + h \frac{qn}{qn-h} + \frac{h^2}{qn-h} + \frac{h}{2(qn-h)} \right) & h > 0, \\ \exp \left(-\beta + h \frac{pn}{pn-h} + \frac{h^2}{pn-h} + \frac{h}{2(pn-h)} \right) & h < 0. \end{cases}$$

Proof. This lemma is an adaptation of the arguments in [15, Chapter 1]. Here we utilize the inequalities, valid for all $0 < t < 1$,

$$\begin{aligned} \frac{-t}{1-t} &< \log(1-t) < -t \\ 0 &< \log(1+t) < t. \end{aligned}$$

The main difference is that we do not place any added restrictions on h other than $-np < h < np$. \square

Lemma 3.17. *Suppose N is a Binomial distribution with parameters n and p , with $\mu := \sup_n \mathbb{E} N < \infty$. Then we have*

$$\mathbb{P}(N \geq \log(n)) < \frac{1}{n}$$

for all $n \geq n_0$, where we may take

$$n_0 = \exp(\mu e^2).$$

Proof. The Hoeffding bound for the binomial distribution, see for example [4, equation (41)], implies that for any $y > 1$, we have

$$\mathbb{P}(N \geq y \mathbb{E} N) \leq y^{-y \mathbb{E} N} e^{(y-1) \mathbb{E} N}.$$

Furthermore, taking $y = \frac{1}{\mathbb{E} N} \log(n)$, we then solve for n in

$$y^{-y \mathbb{E} N} e^{(y-1) \mathbb{E} N} < 1/n;$$

rearranging, and taking the logarithm, we wish to satisfy

$$\log(\mathbb{E} N) + 2 - \frac{\mathbb{E} N}{\log(n)} < \log \log n.$$

Next, we may ignore the term $\frac{\mathbb{E} N}{\log(n)}$ as long as $n > e^{\mathbb{E} N}$, in which case we obtain after exponentiating twice

$$n > e^{e^2 \mathbb{E} N},$$

and replacing $\mathbb{E} N$ with μ we obtain the conclusion. \square

Lemma 3.18. *Let N_3, N_4, \dots be defined as in (38), with p_i given by (36), ρ given by (56), and $x = \frac{2m_1}{m_2} \frac{n-k}{k}$. Then*

$$\mathbb{P}(3N_3 + 4N_4 + \dots \geq \log(n)) \leq \frac{1}{n}$$

for all $n \geq n_3$, where n_3 is the smallest value which satisfies $x \rho \leq \frac{1}{2}$ and

$$(67) \quad \frac{n(n-k)^3}{k^2} \leq \frac{m_1}{2\rho^4} \left(\frac{2m_1}{m_2} \right)^3.$$

Proof. We have

$$\mathbb{P}(3N_3 + 4N_4 + \dots \geq \log(n)) \leq \mathbb{P}(N_3 + N_4 + \dots > 0) \leq \mathbb{E}(N_3 + N_4 + \dots) \leq k \frac{2\rho^4 x^3}{m_1}.$$

By plugging in the appropriate values for x and rearranging, we obtain the result. \square

Theorem 3.19. *Consider an assembly as governed by (10) and (13), and assume further that $m_1, m_2, m_3 > 0$ and $M(\cdot)$ has a strictly positive radius of convergence. Given $n \geq k \geq 1$, pick an assembly uniformly from the $p(n, k)$ choices having exactly k components. Let*

$$x = \frac{2m_1}{m_2} \frac{n-k}{k}, \quad b_0 = m_1 x, \quad b_1 = \frac{m_2 x^2}{2}, \quad b_2 = \frac{m_3 x^3}{6},$$

and define functions

$$f(h, n, p) := -\frac{h}{2pn} + \frac{h}{2qn} - \frac{h^2}{pn} - \frac{h^2}{qn},$$

$$g(h, n, p) := h\frac{pn}{pn-h} + \frac{h^2}{pn-h} + \frac{h}{2(pn-h)}.$$

Finally, define

$$\lambda \equiv \lambda(n, k) := \frac{m_3}{6m_1}x^2, \quad \text{and} \quad p := \frac{b_1}{b_0 + b_1}.$$

Then the number of components of size 3, $D_3(n, k)$, satisfies

$$\mathbb{P}(D_3(n, k) = m) \geq \frac{\lambda^k}{k!} e^{-\lambda} e^{-\frac{k^2}{n-k} - \frac{\lambda^2}{n-\lambda}} \frac{n-1}{n} \frac{\exp(-\beta + g(2m + \log(n), k, p))}{\exp(f(2m + \log(n), k, p))},$$

and

$$\mathbb{P}(D_3(n, k) = m) \leq \frac{\lambda^k}{k!} e^{-\lambda} e^{\frac{\lambda m}{k}} \frac{\frac{1}{n} + \frac{1}{\sqrt{2\pi}r} \exp(f(\log(n), k - \log(n), p))}{\frac{n-1}{n} \frac{1}{\sqrt{2\pi}r} \exp(-\beta + g(\log(n), k - \log(n), p))};$$

and the largest component L_1 satisfies

$$e^\lambda \mathbb{P}(L_1 = 2) \leq \frac{\exp(f(0, k, p))}{\frac{n-1}{n} \exp(-\beta + g(\log(n), k, p))}$$

and

$$e^\lambda \mathbb{P}(L_1 = 2) \geq e^{-\frac{\lambda^2}{n-\lambda}} \frac{\exp(-\beta + g(0, k, p))}{\frac{n-1}{n} \exp(f(\log(n), k, p))} \left(1 - \frac{1}{n(1 - b_2/(m_1 x))}\right)$$

for all $n \geq \max(e^{\mu e^2}, n_3)$, where $\mu = \sup_{n,k} \frac{k b_1}{b_0 + b_1}$, and n_3 is the smallest positive value such that $x\rho \leq \frac{1}{2}$ and satisfies (67).

Proof. We start by defining the events $B_2 := \{2N_2 \geq \frac{1}{2} \log(n)\}$ and $B_3 := \{3N_3 + 4N_4 + \dots \geq \frac{1}{2} \log(n)\}$. By Lemma 3.17, we have $\mathbb{P}(B_2) < \frac{1}{n}$ for all $n > e^{\mu_2 e^2}$, where $\mu_2 := \max \frac{k m_3 x^2}{3!}$. By Lemma 3.18, we have $\mathbb{P}(B_3) < \frac{1}{n}$ for all $n > n_3$, where we may take n_3 to be the smallest positive value such that

$$\frac{n(n-k)^3}{k^2} \leq \frac{m_1}{2\rho^4} \left(\frac{m_2}{2m_1}\right)^3.$$

Combining, the bad event $B := B_2 \cup B_3$ has $\mathbb{P}_x(B) < 2/n$ for all $n \geq \max(n_2, n_3)$, and on the complementary event, B^c , with $k_1 = r_1 := r_2 + r_3 = \log n$, we have $K_0 := N_2 + N_3 + \dots \leq k_1$ and $R_0 := 2N_2 + 3N_3 + \dots \leq r_1$.

Conditional on $\{K_0 = k_0\}$ and further conditioning on *which* of the $k - k_0$ indices $j \in [k]$ did not contribute to K_0 , i.e., those j for which $X_j = 0$ or 1, we have $k - k_0$ independent trials where $p := \mathbb{P}(X_j = 1) = 1 - \mathbb{P}(X_j = 0) = p_1/(p_0 + p_1) = \frac{b_1}{b_0 + b_1}$, which defines the binomial distribution for N_1 conditioned on the values of N_2, N_3, \dots . We thus obtain bounds on $\mathbb{P}(N_1 + 2N_2 + \dots = r)$ by splitting it up by B and B^c . Conditional on B

or B^c , N_1 is binomial with parameters $k - k_0$ and $p = p_1/(p_0 + p_1)$. Let $R(r) := \{N_1 + 2N_2 + \dots = r\}$, and $\beta = \frac{1}{12r} + \frac{1}{12(k-r)}$. Let $f(h, n, p) := -\frac{h}{2pn} + \frac{h}{2qn} - \frac{h^2}{pn} - \frac{h^2}{qn}$, and $g(h, n, p) := h\frac{pn}{pn-h} + \frac{h^2}{pn-h} + \frac{h}{2(pn-h)}$. With $k_0 = \log(n)$, we have

$$\mathbb{P}(R(r) \cap B^c) \leq \mathbb{P}(N_1 = r) \leq \frac{1}{\sqrt{2\pi r}} \exp(f(\log(n), k - k_0, p)),$$

and

$$\begin{aligned} \mathbb{P}(R(r) \cap B^c) &\geq \frac{n-1}{n} \max_{0 \leq \ell \leq \log(n)} \mathbb{P}(N_1 = r - \ell) \\ &\geq \frac{n-1}{n} \mathbb{P}(N_1 = r - r_2 - r_3) \\ &\geq \frac{n-1}{n} \frac{1}{\sqrt{2\pi r}} \exp(-\beta + g(\log(n), k - k_0, p)). \end{aligned}$$

Hence,

$$\frac{n-1}{n} \frac{1}{\sqrt{2\pi r}} \exp(-\beta + g(\log(n), k - \log(n), p)) \leq \mathbb{P}(R(r))$$

and

$$\mathbb{P}(R(r)) \leq \frac{1}{n} + \frac{1}{\sqrt{2\pi r}} \exp(f(\log(n), k - \log(n), p)).$$

Next, we obtain bounds on $\mathbb{P}(N_2 = m | N_1 + 2N_2 + \dots = r)$. We have

$$\begin{aligned} \frac{\mathbb{P}(N_2 = m | R(r))}{\mathbb{P}(N_2 = m)} &\geq \frac{\mathbb{P}(B_3) \min_{0 \leq \ell \leq \log(n)} \mathbb{P}(N_1 = r - 2m - \ell)}{\mathbb{P}(N_1 + 2N_2 + \dots = r)} \\ &\quad + \frac{\mathbb{P}(B_3^c) \min_{0 \leq \ell \leq \log(n)} \mathbb{P}(N_1 = r - 2m - \ell)}{\mathbb{P}(N_1 + 2N_2 + \dots = r)} \\ &> \frac{n-1}{n} \frac{\mathbb{P}(N_1 = r - 2m - \log(n))}{\mathbb{P}(N_1 + 2N_2 + \dots = r)} \\ &> \frac{n-1}{n} \frac{\frac{1}{\sqrt{2\pi r}} \exp(-\beta + g(2m + \log(n), k, p))}{\frac{1}{\sqrt{2\pi r}} \exp(f(2m + \log(n), k, p))}. \end{aligned}$$

In the other direction, we have

$$\begin{aligned} \frac{\mathbb{P}(N_2 = m | R(r))}{\mathbb{P}(N_2 = m)} &< \frac{\frac{1}{n} + \mathbb{P}(N_1 = r - 2m)}{\mathbb{P}(N_1 + 2N_2 + \dots = r)} \\ &\leq \frac{\frac{1}{n} + \frac{1}{\sqrt{2\pi r}} \exp(f(\log(n), k - \log(n), p))}{\frac{n-1}{n} \frac{1}{\sqrt{2\pi r}} \exp(-\beta + g(\log(n), k - \log(n), p))}. \end{aligned}$$

At this point we apply Lemma 3.15 to $\mathbb{P}(N_2 = m)$ to obtain the result.

Let us now consider $\mathbb{P}(L_1 = 2) = \mathbb{P}(N_1 > 0, N_2 = N_3 = \dots = 0 | R(r))$. Let $T_3 = \{N_3 = N_4 = \dots = 0\}$. We have

$$\begin{aligned} \mathbb{P}(L_1 = 2) &= \frac{\mathbb{P}(N_1 > r, N_2 = 0, T_3)}{\mathbb{P}(R(r))} \\ &= \frac{\mathbb{P}(N_1 = r | R(r), N_2, T_3) (1 - p_2)^k \mathbb{P}(T_3 | R(r), N_2 = 0)}{\mathbb{P}(R(r))} \\ &= \frac{\mathbb{P}(\text{Bin}(k, p) = r)}{\mathbb{P}(R(r))} (1 - p_2)^k \left(1 - \frac{p_3 + p_4 + \dots}{1 - p_2}\right)^k. \end{aligned}$$

Whence,

$$\begin{aligned} \mathbb{P}(L_1 = 2) &\leq \mathbb{P}(N_2 = 0) \frac{\frac{1}{\sqrt{2\pi r}} \exp(f(0, k, p))}{\frac{n-1}{n} \frac{1}{\sqrt{2\pi r}} \exp(-\beta + g(\log(n), k, p))} \\ &\leq e^{-\lambda} \frac{\exp(f(0, k, p))}{\frac{n-1}{n} \exp(-\beta + g(\log(n), k, p))}. \end{aligned}$$

In a similar fashion, we have

$$\begin{aligned} \mathbb{P}(L_1 = 2) &\geq \mathbb{P}(N_2 = 0) \frac{\frac{1}{\sqrt{2\pi r}} \exp(-\beta + g(0, k, p))}{\frac{n-1}{n} \frac{1}{\sqrt{2\pi r}} \exp(f(\log(n), k, p))} \left(1 - \frac{p_3 + p_4 + \dots}{1 - p_2}\right)^k \\ &\geq \mathbb{P}(N_2 = 0) \frac{\exp(-\beta + g(0, k, p))}{\frac{n-1}{n} \exp(f(\log(n), k, p))} \left(1 - \frac{k(p_3 + p_4 + \dots)}{1 - p_2}\right) \\ &\geq e^{-\lambda} e^{-\frac{\lambda^2}{k-\lambda}} \frac{\exp(-\beta + g(0, k, p))}{\frac{n-1}{n} \exp(f(\log(n), k, p))} \left(1 - \frac{k(p_3 + p_4 + \dots)}{1 - p_2}\right) \\ &\geq e^{-\lambda} e^{-\frac{\lambda^2}{k-\lambda}} \frac{\exp(-\beta + g(0, k, p))}{\frac{n-1}{n} \exp(f(\log(n), k, p))} \left(1 - \frac{2k\rho^4 x^3}{m_1(1 - p_2)}\right). \end{aligned}$$

Now whenever $n \geq n_3$, we have $\frac{2k\rho^4 x^3}{m_1(1-p_2)} \leq \frac{1}{n}$, and so we have

$$\mathbb{P}(L_1 = 2) \geq e^{-\lambda} e^{-\frac{\lambda^2}{k-\lambda}} \frac{\exp(-\beta + g(0, k, p))}{\frac{n-1}{n} \exp(f(\log(n), k, p))} \left(1 - \frac{1}{n(1 - p_2)}\right).$$

Finally, since $p_2 = b_2/M(x) \leq b_2/(m_1 x)$, we obtain the final expression. \square

4. ACKNOWLEDGEMENTS

The authors are grateful to Fred Kochman for helpful suggestions.

REFERENCES

- [1] ALMKVIST, G. Partitions into odd, unequal parts. *Journal of Pure and Applied Algebra* 38, 2-3 (1985), 121–126. [4](#)
- [2] ARRATIA, R., BARBOUR, A. D., AND TAVARÉ, S. On random polynomials over finite fields. *Mathematical Proceedings of the Cambridge Philosophical Society* 114, 2 (1993), 347–368. [11](#)

- [3] ARRATIA, R., BARBOUR, A. D., AND TAVARÉ, S. *Logarithmic combinatorial structures: a probabilistic approach*. EMS Monographs in Mathematics. European Mathematical Society (EMS), Zürich, 2003. 6, 8, 11
- [4] ARRATIA, R., AND BAXENDALE, P. Bounded size bias coupling: a Gamma function bound, and universal Dickman-function behavior. *Probability Theory and Related Fields* 162, 3-4 (2015), 411–429. 20, 29
- [5] ARRATIA, R., AND DESALVO, S. Approximation to the component sizes of low-rank set partitions and permutations, via rooks. *Preprint*. 2
- [6] ARRATIA, R., AND DESALVO, S. Completely effective error bounds for Stirling numbers of the first and second kind via Poisson approximation. *Annals of Combinatorics* (2016). 2, 4
- [7] ARRATIA, R., AND DESALVO, S. Probabilistic divide-and-conquer: A new exact simulation method, with integer partitions as an example. *Combinatorics, Probability and Computing* 25 (5 2016), 324–351. 15
- [8] ARRATIA, R., AND TAVARÉ, S. The cycle structure of random permutations. *The Annals of Probability* 20, 3 (1992), 1567–1591. 11
- [9] ARRATIA, R., AND TAVARÉ, S. Independent process approximations for random combinatorial structures. *Advances in Mathematics* 104, 1 (1994), 90–154. Available at <http://arxiv.org/pdf/1308.3279.pdf>. 2, 6, 14, 15
- [10] BENDER, E., AND GOLDMAN, J. Enumerative uses of generating functions. *Indiana University Mathematics Journal* 20 (1971), 753–765. 8
- [11] BENDER, E. A. Central and local limit theorems applied to asymptotic enumeration. *Journal of Combinatorial Theory. Series A* 15 (1973), 91–111. 4
- [12] BENDER, E. A., AND RICHMOND, L. B. Central and local limit theorems applied to asymptotic enumeration. II. Multivariate generating functions. *Journal of Combinatorial Theory. Series A* 34, 3 (1983), 255–265. 4
- [13] BENDER, E. A., RICHMOND, L. B., AND WILLIAMSON, S. G. Central and local limit theorems applied to asymptotic enumeration. III. Matrix recursions. *Journal of Combinatorial Theory. Series A* 35, 3 (1983), 263–278. 4
- [14] BERLEKAMP, E. R. *Algebraic coding theory*. McGraw-Hill Book Co., New York-Toronto, Ont.-London, 1968. 8
- [15] BOLLOBÁS, B. *Random graphs*, second ed., vol. 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, 2001. 28
- [16] BRENTI, F. Log-concave and unimodal sequences in algebra, combinatorics, and geometry: an update. In *Jerusalem combinatorics '93*, vol. 178 of *Contemp. Math.* Amer. Math. Soc., Providence, RI, 1994, pp. 71–89. 4
- [17] CANFIELD, E. R. From recursions to asymptotics: on Szekeres' formula for the number of partitions. *Electronic Journal of Combinatorics* 4, 2 (1997), Research Paper 6, approx. 16 pp. (electronic). The Wilf Festschrift (Philadelphia, PA, 1996). 4
- [18] CHELLURI, R., RICHMOND, L. B., AND TEMME, N. M. Asymptotic estimates for generalized Stirling numbers. *Analysis. International Mathematical Journal of Analysis and its Applications* 20, 1 (2000), 1–13. 4
- [19] DANIELS, H. E. Saddlepoint approximations in statistics. *Annals of Mathematical Statistics* 25 (1954), 631–650. 6
- [20] DESALVO, S. Probabilistic divide-and-conquer: deterministic second half. *arXiv preprint arXiv:1411.6698* (2014). 15
- [21] DESALVO, S., AND PAK, I. Log-concavity of the partition function. *The Ramanujan Journal* (10 2013). 4, 5
- [22] DUCHON, P., FLAJOLET, P., LOUCHARD, G., AND SCHAEFFER, G. Boltzmann samplers for the random generation of combinatorial structures. *Combinatorics, Probability and Computing* 13, 4-5 (2004), 577–625. 14

- [23] ERDOS, P., AND LEHNER, J. The distribution of the number of summands in the partitions of a positive integer. *Duke Mathematics Journal* 8, 2 (1941), 335–345. [3](#), [17](#)
- [24] ESSCHER, F. On the probability function in the collective theory of risk. *Skandinavisk Aktuarietidskrift* 15 (1932), 175–195. [6](#)
- [25] FELLER, W. *An Introduction to Probability Theory and Its Applications. Vol. I*. John Wiley & Sons, Inc., New York, N.Y., 1950. [28](#)
- [26] FLAJOLET, P., AND SEDGEWICK, R. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009. [8](#)
- [27] FOATA, D. *La série génératrice exponentielle dans les problèmes d'énumération*. Les Presses de l'Université de Montréal, Montreal, Que., 1974. Avec un chapitre sur les identités probabilistes dérivées de la formule exponentielle, par B. Kittel, Séminaire de Mathématiques Supérieures, No. 54 (Été, 1971). [8](#)
- [28] FOATA, D., AND SCHÜTZENBERGER, M. P. *Théorie géométrique des polynômes eulériens*. Lecture Notes in Mathematics, Vol. 138. Springer-Verlag, Berlin-New York, 1970. [8](#)
- [29] GRANVILLE, A. Cycle lengths in a permutation are typically Poisson. *Electron. J. Combin.* 13, 1 (2006), Research Paper 107, 23. [2](#)
- [30] HARDY, G. H., AND RAMANUJAN, S. Asymptotic formulae in combinatory analysis. *Proceedings of the London Mathematical Society* (1918), 75 – 115. [4](#), [8](#)
- [31] JENSEN, J. L. *Saddlepoint approximations*, vol. 16 of *Oxford Statistical Science Series*. The Clarendon Press, Oxford University Press, New York, 1995. Oxford Science Publications. [6](#)
- [32] JORDAN, C. *Calculus of finite differences*. Third Edition. Introduction by Harry C. Carver. Chelsea Publishing Co., New York, 1965. [4](#)
- [33] JOYAL, A. Une théorie combinatoire des séries formelles. *Advances in Mathematics* 42, 1 (1981), 1–82. [8](#)
- [34] JOYAL, A. Foncteurs analytiques et espèces de structures. In *Combinatoire énumérative (Montreal, Que., 1985/Quebec, Que., 1985)*, vol. 1234 of *Lecture Notes in Math*. Springer, Berlin, 1986, pp. 126–159. [8](#)
- [35] LEHMER, D. H. On the series for the partition function. *Transactions of the American Mathematical Society* 43, 2 (1938), 271–295. [8](#)
- [36] LEHMER, D. H. On the remainders and convergence of the series for the partition function. *Transactions of the American Mathematical Society* 46 (1939), 362–373. [5](#), [8](#)
- [37] LOUCHARD, G. Asymptotics of the Stirling numbers of the first kind revisited: a saddle point approach. *Discrete Mathematics & Theoretical Computer Science. DMTCS*. 12, 2 (2010), 167–184. [4](#)
- [38] LOUCHARD, G. Asymptotics of the Stirling numbers of the second kind revisited. *Applicable Analysis and Discrete Mathematics* 7, 2 (2013), 193–210. [4](#)
- [39] MOSER, L., AND WYMAN, M. Stirling numbers of the second kind. *Duke Mathematical Journal* 25 (1957), 29–43. [4](#)
- [40] MOSER, L., AND WYMAN, M. Asymptotic development of the Stirling numbers of the first kind. *Journal of the London Mathematical Society. Second Series* 33 (1958), 133–146. [4](#)
- [41] NICOLAS, J.-L. Sur les entiers N pour lesquels il y a beaucoup de groupes abéliens d'ordre N . *Université de Grenoble. Annales de l'Institut Fourier* 28, 4 (1978), 1–16, ix. [5](#)
- [42] PITTEL, B. On a likely shape of the random Ferrers diagram. *Advances in Applied Mathematics* 18, 4 (1997), 432–488. [6](#)
- [43] PITTEL, B. Random set partitions: asymptotics of subset counts. *Journal of Combinatorial Theory. Series A* 79, 2 (1997), 326–359. [6](#)

- [44] RADEMACHER, H. A convergent series for the partition function $p(n)$. *Proceedings of the National Academy of Sciences* 23 (1937), 78–84. 4, 8
- [45] REID, N. Saddlepoint methods and statistical inference. *Statistical Science. A Review Journal of the Institute of Mathematical Statistics* 3, 2 (1988), 213–238. With comments and a rejoinder by the author. 6
- [46] RIDDELL, JR., R. J., AND UHLENBECK, G. E. On the theory of virial development of the equation of state of monoatomic gases. *The Journal of Chemical Physics* 21 (1953), 2056–2064. 8
- [47] ROMIK, D. Partitions of n into $t\sqrt{n}$ parts. *European Journal of Combinatorics* 26, 1 (2005), 1–17. 4
- [48] ROSSER, J. B., AND SCHOENFELD, L. Approximate formulas for some functions of prime numbers. *Illinois J. Math.* 6 (1962), 64–94. 3
- [49] SHEPP, L. A., AND LLOYD, S. P. Ordered cycle lengths in a random permutation. *Transactions of the American Mathematical Society* 121 (1966), 340–357. 11
- [50] STANLEY, R. P. Log-concave and unimodal sequences in algebra, combinatorics, and geometry. In *Graph theory and its applications: East and West (Jinan, 1986)*, vol. 576 of *Ann. New York Acad. Sci.* New York Acad. Sci., New York, 1989, pp. 500–535. 4
- [51] SZEKERES, G. An asymptotic formula in the theory of partitions. *The Quarterly Journal of Mathematics. Oxford. Second Series* 2 (1951), 85–108. 3
- [52] SZEKERES, G. Some asymptotic formulae in the theory of partitions. II. *The Quarterly Journal of Mathematics. Oxford. Second Series* 4 (1953), 96–111. 3, 4
- [53] VAN LINT, J. H., AND WILSON, R. M. *A course in combinatorics*, second ed. Cambridge University Press, Cambridge, 2001. 8
- [54] WILF, H. S. *generatingfunctionology*. Academic Press, Inc., Boston, MA, 1990. 8
- [55] WILLIAMS, D. *Diffusions, Markov processes, and martingales. Vol. 1*, 2 ed. John Wiley & Sons, Ltd., Chichester, 1979. Foundations, Probability and Mathematical Statistics. 12